Escuela Politécnica Superior

Departamento de Tecnología Electrónica y de las Comunicaciones

OBJECT DETECTION FOR VIDEO-MONITORING USING FIXED MULTI-CAMERA SYSTEMS

PhD Thesis written by
Rafael Martín Nieto
under the supervision of
Dr. José María Martínez Sánchez
and
Dr. Álvaro García Martín

Madrid, May 2018

Department: Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid, Spain

PhD Thesis: Object detection for video-monitoring using fixed multi-camera systems.

Author: **Rafael Martín Nieto**
Ingeniero de Telecomunicación
Universidad Autónoma de Madrid , Spain

Supervisor: **Jose María Martínez Sánchez**
Doctor Ingeniero de Telecomunicación
Universidad Autónoma de Madrid , Spain

Supervisor: **Álvaro García Martín**
Doctor Ingeniero de Telecomunicación
Universidad Autónoma de Madrid , Spain

Year: 2018

Comittee: _____

_____

_____

The work described in this Thesis was carried out within the Video Processing and Understanding Lab at the Department of Tecnología Electrónica y de las Comunicaciones, Escuela Politécnica Superior, Universidad Autónoma de Madrid (from 2014 to 2018). It was partially supported by the Spanish Government (TEC2014-53176-R, HAVideo) and by the Spanish Government FPU grant programme (Ministerio de Educación, Cultura y Deporte).

To my family and friends.

To win big, you sometimes have to take big risks.
- Bill Gates

Acknowledgments

First of all, I would like to express my gratitude to my supervisor, Dr. José M. Martínez. From the first year of university he gave me the opportunity to work with the laboratory, until today, and dedicating time to me year after year, despite his busy schedule. I also want to thank Dr. Jesús Bescós for his close and personal treatment. Thanks to all present and past members of VPU-Lab, especially Álvaro, who has been a friend as well as a tutor. I also want to thank the people who have helped with the writing and validation procedures of the thesis: to the department reviewer Rubén Vera and to the external reviewers Dr. Alex Hauptmann and Kevin McGuinness. I would also like to thank Dr. Alex Hauptmann from the Carnegie Mellon University for their support, advice and warm hospitality during my short stay.

Quiero agradecer a Carolina de Santiago, que me orientó y me ayudó de elegir mi rama académica, y que sin duda es la que más influyó en la trayectoria profesional que ha acabado en esta tesis. También agradecer a mi familia por darme un hogar, cariño y estabilidad para poder dedicar estos años al doctorado. No puedo olvidarme de Javi, mi compañero de batallas de todos estos años, y a todos mis amigos: los kiki's, los monsters, los tiburones, los supervivientes, los 100% y todos los demás.

Pido disculpas a todos los que haya olvidado, y agradezco a todas las personas a las que aprecio y con las que he tratado estos años, ya que gracias a ellos soy quien soy hoy por hoy.

Rafael Martín Nieto
May 2018

Abstract

Object detection is one of the most important tasks in computer vision. This is a very complex task due to the difficulty of modelling objects, which contains a high degree of variability, and its performance is also very dependent on the data used for training. There are multiple detection algorithms in the state of the art, but all them present problems with one or multiple factors like: occlusions, illumination changes, perspective changes, etc. This thesis addresses tasks related to object detection: training and evaluation framework, detection approaches and applications, and detection improvements in multi-camera scenarios. In the first part of this thesis, we focus on the training and evaluation framework. We analyze existing datasets in the state of the art that meet the requirements we need to evaluate the different developed systems. These datasets must be multi-camera datasets, in which the cameras have an orientation that generate overlap between the points of view. To complete these existing datasets, two new datasets have been designed, recorded and published: one containing wheelchair users, and another one which contains vehicles in a parking lot. Continuing with the evaluation framework, we present the metrics commonly used for the evaluation of object detectors. First, the "classical" evaluation metrics are formulated, precision and recall, and their combinations. For the evaluation of some of the different developed applications, we adapt these metrics for, on the one hand, considering a third dimension (depth) in the scenarios, and, on the other hand, evaluating the capacity to detect occupied or empty parking spots. To finish this part, we present a technique for the generation of synthetic datasets to be able to train a detection model without having enough training data. We train a wheelchair user model considering synthetic datasets from empty wheelchairs images and standing people images. Three synthetic image datasets have been created in order to train three different models, evaluating which model is optimal, and, finally, analyzing its feasibility by comparing it with a people detector model for wheelchair users trained with real images. In the second part, this thesis presents two different detection approaches with a final application. With the idea of providing an existing object detector model the capacity to detect variants of the desired object, which have not been considered in their initial design, we present a wheelchair users model and we include it in a generic people detector, providing a more general solution to detect people in environments such as houses adapted for independent and assisted living, hospitals, healthcare centers and senior residences. As an application of the

presented work, an example of a room in a nursing home is shown in which the detections are mapped on the ground plane in order to monitor people. To conclude this part, we present an automatic multi-camera system for vehicle detection and their corresponding mapping into the parking spots of a parking lot. The results clearly show that the proposed system works correctly in challenging scenarios including almost total occlusions, illumination changes and different weather conditions. Finally, the third part of the thesis takes as starting point the output of the detection algorithms executed on the images and sequences, adding performance improvements and autoparameterization of the algorithms. We combine information obtained from different cameras in order to enhance object detection algorithms performance. Using multiple cameras and information from the recorded scenario, called contextual information (distances between detected objects and cameras, position of the cameras, etc.), the detection performance is improved taking advantage of the results of the other cameras, transferring information from one camera to another, and then combining it. This technique also allows, using an additional correlation framework, to automatically adapt (determining an optimal threshold for each camera) and improve any detector in multi-camera scenarios, during runtime detection.

Resumen

La detección de objetos es una de las principales tareas de visión por ordenador. Esta tarea tiene una gran complejidad debido a la dificultad para modelar objetos, ya que estos contienen un alto grado de variabilidad y su rendimiento es además muy dependiente de los datos usados para su entrenamiento. Hay múltiples algoritmos de detección en el estado del arte, pero todos ellos presentan problemas con uno o varios factores tales como: oclusiones, cambios de iluminación, cambios de perspectiva, etc. Esta tesis aborda tareas relacionadas con la detección de objetos: el marco de entrenamiento y evaluación, aproximaciones y aplicaciones de detección, y mejoras de detección en escenarios multi-cámara. En la primera parte de esta tesis, nos centramos en el entrenamiento y marco de evaluación. Analizamos los conjuntos de datos existentes en el estado del arte que cumplen los requisitos que necesitamos para evaluar los distintos sistemas desarrollados. Estos conjuntos de datos deben ser multi-cámara, en los que las cámaras poseen una orientación que genera solapamiento entre los puntos de vista. Para completar estos conjuntos de datos, se han diseñado, grabado y publicado dos nuevos conjuntos de datos: el primero contiene usuarios de sillas de ruedas, y el segundo contiene vehículos en un parking. Continuando con el marco de evaluación, presentamos las métricas usadas comúnmente para la evaluación de detectores de objetos. Primero se formulan las métricas de evaluación 'clásicas', precisión y exhaustividad, y sus combinaciones. Para la evaluación de algunas de las distintas aplicaciones desarrolladas, adaptamos estas métricas para, por un lado, considerar una tercera dimensión (profundidad) en los escenarios y, por otro lado, evaluar la capacidad de detectar plazas de aparcamiento ocupadas y vacías. Para terminar esta parte, presentamos una técnica para la generación de conjuntos de entrenamiento sintéticos, que permiten entrenar un modelo de detección en situaciones en las que no se dispone de suficientes datos de entrenamiento. Se ha entrenado un modelo de usuario de sillas de ruedas considerando conjuntos de datos sintéticos de sillas de ruedas desocupadas y personas de pie. Se han creado tres conjuntos de datos sintéticos con el fin de entrenar tres modelos distintos, evaluando qué modelo es más óptimo y, finalmente, analizando su viabilidad comparándolos con un modelo de detector de personas para usuarios de sillas de ruedas entrenado con imágenes reales. En la segunda parte, esta tesis presenta dos aproximaciones de detección de objetos, con aplicación final. Con la idea de proveer a un detector de objetos existente con la capacidad de detectar variantes del objeto deseado, las cuales no han sido consideradas en

su diseño inicial, presentamos un modelo de persona en silla de ruedas y lo incluimos en un detector de personas genérico, obteniendo una solución más general para detectar personas en entornos tales como casas adaptadas para la vida independiente y asistida, hospitales, centros de salud y residencias de ancianos. Como aplicación del trabajo presentado, se muestra un ejemplo de una sala de una residencia de ancianos en la que las detecciones se mapean en el plano del suelo con el fin de monitorizar a las personas. Para concluir esta parte, presentamos un sistema automático multi-cámara para detección de vehículos y su correspondiente mapeo en las plazas de aparcamiento de un parking. Los resultados claramente muestran que el sistema propuesto funciona correctamente en escenarios que presentan dificultades como oclusiones casi totales, cambios de iluminación y diferentes condiciones climáticas. Finalmente, la tercera parte de esta tesis toma como punto de partida la salida de los algoritmos de detección ejecutados en las imágenes y secuencias, añadiendo mejoras de rendimiento y autoparametrización de algoritmos, combinando información obtenida de las distintas cámaras con el fin de mejorar el rendimiento de los algoritmos de detección de objetos. Mediante el uso de múltiples cámaras e información del escenario grabado, llamada información contextual (distancia entre los objetos detectados y las cámaras, posición de las cámaras, etc.), el rendimiento de las detecciones se mejora, aprovechando los resultados de las otras cámaras, transfiriendo información de unas cámaras a otras, y después combinando las detecciones. Esta técnica además permite, usando un marco de correlación adicional, adaptar automáticamente (definiendo un umbral óptimo para cada cámara) y mejorando cualquier detector en escenarios multi-cámara, durante el tiempo de ejecución.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction

1.1 Motivation

Computer Vision is a field whose goal is to automate image processing to understand its content. Computer Vision tries to imitate the human vision system in which the brain processes images captured by the eyes. The data may have different formats such as video sequences, different views from multiple cameras and/or multi-modality data. This information is used to solve specific tasks or to understand what happens in the scene.

Object detection is one of the most important tasks in computer vision. This is a very complex task due to the difficulty of modelling objects, which may contain very diverse appearance, pose, attire, point of view, illumination, etc. Real-world scenarios increase the complexity, such as airports, nursing homes, etc., which include multiple people, occlusions and background variability.

This thesis is composed of 3 main parts, and the motivation of each of them is described below.

The motivation of the first part is to achieve a complete training and a rigorous evaluation framework to objectively evaluate the results of each work. The evaluation framework consists of datasets, metrics and detection algorithms. A good set of images and sequences is useful both to train detection models, and to evaluate the results of these models and the systems in which they are integrated. In addition, it is necessary to have enough data to be able to split it into a training set and a test set. Having a complete set of metrics allows evaluating the modified techniques and developed systems, to know if the results have been improved, or if on the contrary, it is necessary to rethink the decisions taken. Finally, the feasibility of training a detection model for which enough training images are not available is studied, which would bring great potential to train a wide range of models.

The second part is motivated by trying to process the results obtained by the object detectors, either by providing an existing detection model the capacity to detect variants of the desired

object, which have not been considered in their initial design, or by video processing, mapping and getting the occupation of a video-monitored parking lot. We generate a wheelchair users dataset to train people detectors, and we add the wheelchair user appearance and poses to a generic people detector. In health care centers, senior residences, hospitals, etc., it is usual to see people who need wheelchairs and their detection is useful to monitor them and to provide them assistance in case they need. Knowing the location of a wheelchair user can be useful for some healthcare applications (e.g. monitoring), and it can be used to analyze the behaviour and actions of such users in different environments. We also consider a second final application for the object detectors, a parking lot management system. The management of the car parks is very expensive and in many cases complex, especially in the case of those that have many places such as airports or large commercial areas. Solving this problem using computer vision promises a number of advantages over intrusive sensors like induction loops or other weight-in-motion sensors. In addition, a vision-based system may provide many value-added services, like parking space guidance and video surveillance. Such systems allow the decongestion of crowded parking areas, directing vehicles to areas with lower occupancy, guiding the vehicles by a faster route.

The motivation of the third part is to achieve a post-processing detection improvement, additional functionality and autoparametrization. As multiple cameras are recording from different points of view, the task to be achieved is to improve the detections of each camera by combining the information obtained by other cameras, without modifying the detection algorithm. Based on this technique, it is also possible to auto-calibrate the parameters of the algorithm without the need, from the person deploying the system, to decide these parameters in a heuristic way.

1.2 Objectives

The main objective of this thesis is to improve video object detection in fixed multi-camera scenarios. For achieving this objective, we propose to focus the three following objectives:

- Training and Evaluation Frameworks objectives:
  - A rigorous, objective and complete evaluation will allow to improve existing systems and methods as the impact of each modification or contribution can be analyzed. We compile and adapt the existing metrics for object detecion evaluation. We adapt the evaluation metrics so that they can be applied to specific applications or so that they can have additional considerations to the basic ones (3D instead of 2D).
  - The creation of new recorded datasets allows to train new detection models, to contemplate new models of appearance, to obtain detection models for specific applications, and to evaluate new techniques or systems. We design, record and publish two new datasets: one for wheelchair users and one for vehicles placed in a parking lot.

  - Training a detection model without having images of the desired object is a problem that involves an investment of time and resources to solve it. We develop a technique for generating synthetic image datasets that can be used to train detection models in situations where a sufficiently large image set of the desired object is not available.

- Detection Approaches and Applications objectives:

  - Contemplating non-conventional appearances for object or people detection allows to obtain more generic detectors or improve the existing ones. We develop a technique to consider the wheelchair users detection in the consideration of this appearance in a generic people detector, and we present an application for video monitoring people in a nursing home.

  - The management of parking lots is very expensive and in many cases complex, especially in the case of those that have many places such as airports or large commercial areas. We propose an automatic system using computer vision which promises a number of advantages over intrusive sensors.

- Detection Improvements in Multi-camera Scenarios objectives:

  - There are multiple detection algorithms in the state of the art, but all them present problems with one or multiple factors like: occlusions, illumination changes, perspective changes, etc. We improve the detection performance taking advantage of the results of other cameras recording the same area from a different point of view.

  - Finding optimal parametrizations for people detectors is a complicated task due to the large number of parameters and the high variability of application scenarios. We develop a framework to automatically adapt and improve any detector in multi-camera scenarios where people are observed from different viewpoints.

1.3 Major contributions

The main contributions of this thesis are summarized below:

1. We analyze and complete the evaluation framework for mono-camera and multi-camera systems focused on object (especially people and vehicles) detection. Two new datasets have been designed, recorded and published.

2. We propose a technique for generating synthetic image datasets and study the feasibility of training an object detection algorithm using synthetic images datasets.

3. We incorporate wheelchair user models for traditional people detectors (we define these as standing people detectors) in order to contemplate this people particular appearance and to be able to detect people globally and in a more generic way.

4. We have designed, implemented and evaluated a multi-camera system for vehicles detection and their corresponding mapping into the parking spots of a parking lot.

5. We improve the detection performance of object detectors using multiple cameras and contextual information from the recorded scenario.

6. We automatically adapt (automatic parameterization) and improve any people detector in multi-camera scenarios where people are observed from various viewpoints.

1.4 Structure of the document

This document is structured in five parts, which are organized as follows:

- Part I: Introduction
  - *Chapter 1: Introduction.* This chapter presents the motivation, the objectives, the main contributions and the structure of this thesis.
- Part II: Training and Evaluation Frameworks
  - *Chapter 2: Existing and proposed datasets, metrics and detection algorithms.* Describes the existing datasets, metrics and detection algorithms, which meet the requirements to evaluate the developed systems and techniques, and presents new datasets and metrics in order to complete the evaluation framework.
  - *Chapter 3: Generation and evaluation of synthetic models for training people detectors.* Various synthetic image datasets have been created in order to train different detection models, evaluating which model is optimal.
- Part III: Detection Approaches and Applications
  - *Chapter 4: Incorporating wheelchair users in people detection.* A wheelchair users detector is presented to extend people detection, providing a more general solution to detect people in environments such as houses adapted for independent and assisted living, hospitals, healthcare centers and senior residences.
  - *Chapter 5: Automatic vacant parking places management system using multi-camera vehicle detection.* An automatic multi-camera system for vehicles detection and their corresponding mapping into the parking spots of a parking lot.

- Part IV: Detection Improvements in Multi-camera Scenarios
  - *Chapter 6: Improving multi-camera people detection using contextual information.* Using multiple cameras and contextual information, the detection performance is improved taking advantage of the results of the other cameras.
  - *Chapter 7: Enhancing multi-camera people detection by stand-alone automatic parametrization using detection transfer and self-correlation maximization.* A framework to automatically adapt and improve any detector in multi-camera scenarios where people are observed from various viewpoints.
- Part V: Conclusions
  - *Chapter 8: Achievements, conclusions and future work.* It concludes this document summarizing the main results and future work for its extension.
- Part VI: Appendixes
  - *Appendix A: Publications.*
  - *Appendix B: Spanish translation of achievements, conclusions and future work.*
- Glossary
- Bibliography

The relationships between chapters and parts of the thesis are depicted in Fig. 1.1.

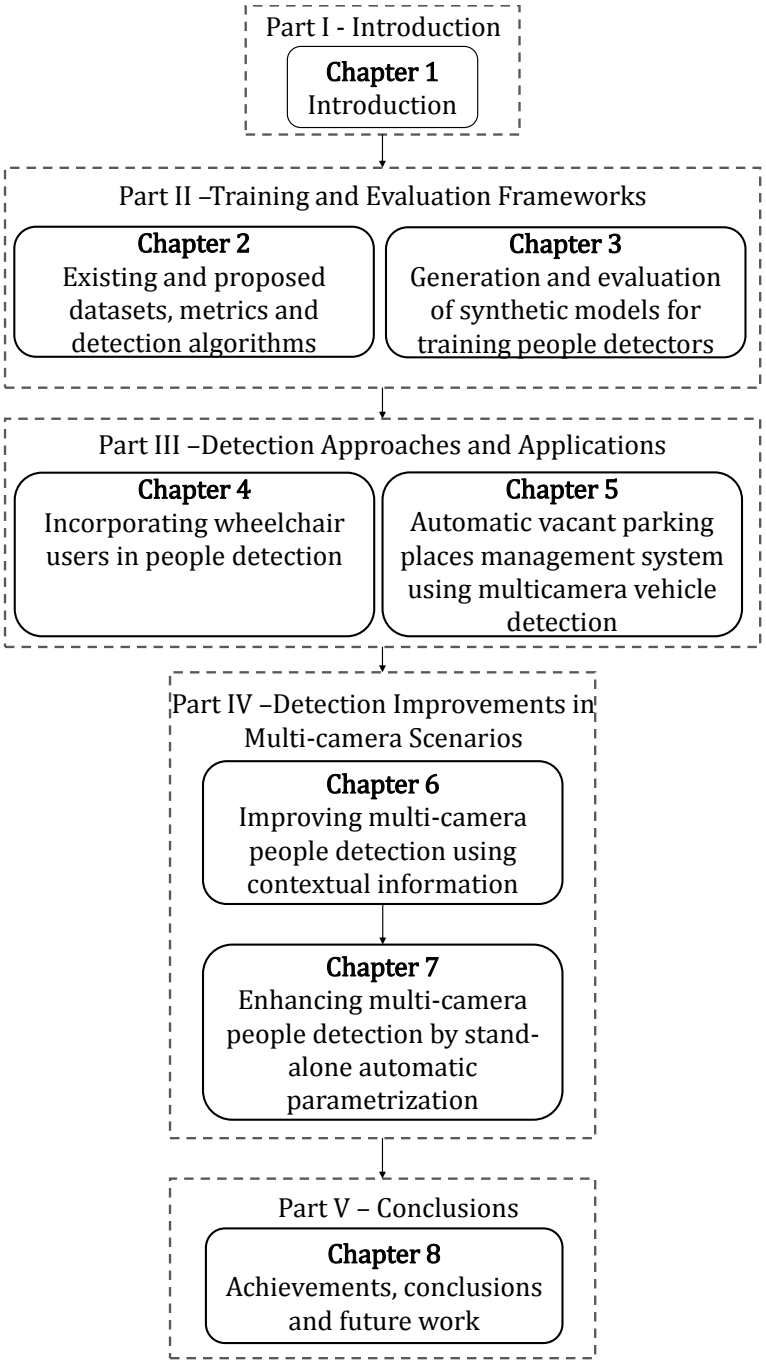


Fig. 1.1. Dependence between the chapters of this thesis.

Part II

Training and Evaluation Frameworks

Chapter 2

Existing and proposed datasets, metrics and detection algorithms

2.1 Introduction

In order to carry out a proper evaluation of the developed systems and algorithms, it is necessary to have a rigorous and complete evaluation framework that allows the evaluation of the capabilities and limitations of the developed works.

First, we focus on analyzing existing datasets in the state of the art that meet the characteristics we need to evaluate the different developed systems. These datasets must be multi-camera datasets, in which the cameras are oriented to generate overlap between the points of view. In addition, objects (usually people) must appear in the recorded sequences to be able to apply the detection algorithms. With these restrictions, there are two existing datasets that meet the restrictions described: PETS2009 and EPFL-RLC. In addition, the use of the SMILE Wheelchair dataset is considered as it contains sequences in which, in addition to standing people, people appear using wheelchairs, which allows to generate detection models for this specific people appearance and also allows to evaluate the work proposed in Chapter 4. To complete these existing datasets, two new datasets have been designed, recorded and published. The first, Wheelchair Users Dataset, is similar to the SMILE Wheelchairs dataset, adding the overlapping multi-camera feature that is required for evaluating the multi-camera contributions in this Thesis. The other recorded dataset, named Parking Lot Dataset, presents a real multi-camera scenario of a vehicle parking, in order to evaluate the system described in Chapter 5, in which instead of detecting people, vehicles are detected in order to obtain the occupancy of a parking lot.

Continuing with the evaluation framework, the chapter presents the metrics commonly used for the evaluation of object detectors. First, the "classical" evaluation metrics are formulated,

precision and recall, and their combinations, the PR-Curve and the F-score. For the evaluation of some of the different developed applications, we adapt these metrics for, on the one hand, considering a third dimension (depth) in the scenarios and, on the other hand, evaluating the capacity to detect occupied or empty parking spots.

This chapter is organized as follows: Section 2.2 presents the considered datasets from the State of the Art, Section 2.3 describes the own designed and recorded datasets, Section 2.5 describe the metrics commonly used for the evaluation of detections, and Section 2.6 describes the evaluation metrics adapted to the specific scenarios or applications developed in this Thesis. Finally, Section 2.8 presents some conclusions.

2.2 Existing datasets

2.2.1 PETS2009

PETS 2009 Benchmark sequences are multisensor sequences containing different crowd activities (http://www.cvg.reading.ac.uk/PETS2009/a.html). This dataset contains outdoor sequences from a typical surveillance setup. The aim of this dataset is to employ existing or new systems for the detection of surveillance characteristics/events within a real-world environment. The cameras are calibrated using Tsai calibration [Tsai, 1986] and the calibration files are included in the dataset.

We also consider the available ground truth from [Milan et al., 2014] which completes the utility of the dataset. View 1 which is the camera that has ground truth available, and over region R1 (see website for details), defined by the dataset owners for the use of multiple views. In addition to view 1, views 5, 6, 7 and 8 are also used in our experiments. The cameras locations and approximate orientation are shown in a satellite map in Figure 2.1, and an example of each camera viewpoint is presented in Figure 2.2. View 8 is facing view 1. Views 5 and 7 are (almost) orthogonal to view 1. Finally, view 6 presents the same orientation than view 1 but at a different distance from the monitored area. The only sequences that have these five points of view available are S2.L1 and S3.MF.

More details and information are available on the dataset website, cited at the beginning of this subsection.

2.2.2 EPFL-RLC Dataset

The EPFL-RLC dataset (https://cvlab.epfl.ch/data/rlc) was recorded in the EPFL Rolex Learning Center using three static HD cameras. Each camera has an original resolution of 1920x1080 pixels but the publicly available frames have a reduced resolution of 480×270 pixels. A frame rate of 60 frames per second was used for each sequence. The cameras are calibrated using Tsai calibration [Tsai, 1986] and the calibration files are included in the dataset. The

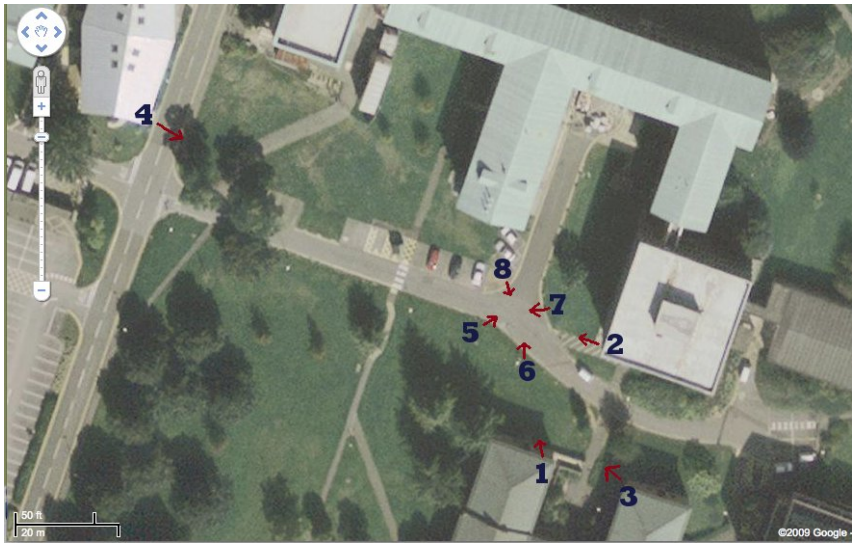


Fig. 2.1. Top view map of the PETS2009 Benchmark (extracted from the dataset website).

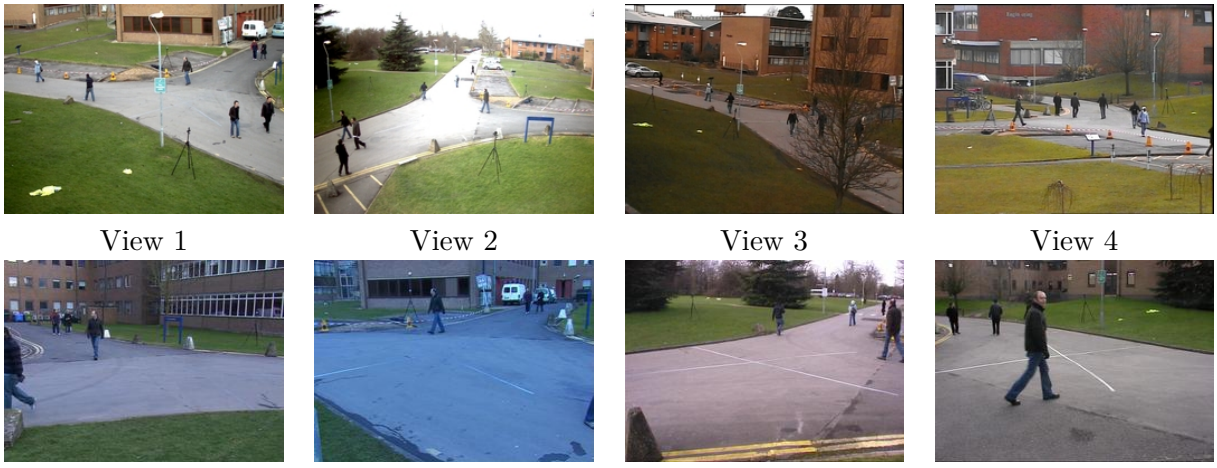


Fig. 2.2. Frame examples of the PETS2009 Benchmark (extracted from the dataset website).

video sequences which are available for download are synchronized across the views and each sequence contains 8000 frames.

The three points of view of the dataset are similar, so we chose camera 1 to generate the ground truth. The ground truth of the frames was not fully annotated so we manually annotated the bounding boxes of the people detections for the first 2000 frames of camera 1. This ground truth is generated to be able to evaluate this dataset in the same way that the other considered datasets of the thesis. We make this ground truth publicly available upon request.

An example of each camera viewpoint is presented in Figure 2.3.

Camera 1 Camera 2 Camera 3

Fig. 2.3. Frame examples of the EPFL-RLC Dataset (extracted from the dataset website).

2.2.3 Smile Wheelchair Dataset

This dataset[1] was created by the Smile Lab at the Department of Electrical Engineering, National Cheng Kung University, Taiwan. The dataset is divided into two main image sets: the train sequences and the test sequences. Each of the frames has a resolution of 720×480 pixels.

The training sequences are composed of 8 image subsets and a total of 3674 images, each one of them contains a set of images of wheelchairs with a defined orientation relative to the camera. The different orientations and models are shown and defined in [Huang et al., 2010].

The test sequences are composed of 4 image subsets, each one of them containing a sequence with a wheelchair and some standing people walking around. Unlike the training set, each of these frame subsets contains a continuous recording, allowing to use tracking techniques to improve detection, as shown in [Huang et al., 2010]. The test set contains a total of 1314 frames divided in 4 groups. Table 2.1 shows the properties of each sequence. More information is provided for this dataset than for the other ones due to it being the only one that is not publicly available.

The ground truth of this dataset was not provided, so we created it annotating manually each of the frames from both sets, training and testing. This ground truth is available for downloading as additional content in the Wheelchair users dataset webpage (http://www-vpu.eps.uam.es/DS/WUds/).

[1]The dataset was courtesy of Smile Lab (http://smile.ee.ncku.edu.tw/) at the Department of Electrical Engineering, National Cheng Kung University, Taiwan.

Sequence number	#Frames	#Wheelchair users	#Standing people
1	449	1	From 3 to 5
2	351	1	From 2 to 5
3	239	1	From 4 to 5
4	287	1	From 3 to 7

Table 2.1: Properties of each of the test sequences from the SMILE Wheelchair dataset.

Sequence 1 Sequence 2 Sequence 3 Sequence 4

Fig. 2.4. Frame examples of the SMILE Wheelchair Dataset (extracted from the dataset).

2.3 Proposed datasets

2.3.1 Wheelchair Users dataset

This dataset was recorded by the Video Processing and Understanding Lab due to the lack of public wheelchair datasets. We used it to test the trained wheelchair users detector, as it contains sequences with a higher number of wheelchairs (up to four) and some more complex situations and scenarios (illumination changes, occlusions, etc.). The sequences were recorded in a real environment of a senior residence, in order to work with an environment as realistic as possible.Due to privacy issues, real recording with actual residents was not possible, so we recorded sequences with people acting as wheelchair users. Each of the frames has a resolution of 768×432 pixels and the sequences are recorded at 25 fps. Compared to the other wheelchair dataset, this one contains a new environment with a larger number of sequences, a greater number of frames per sequence, and more wheelchair types (three different wheelchairs).

The dataset consists of 11 sequences (S1 to S11), each of them recorded from two points of views (V1 and V2), resulting in a total of 22 sequences. Table 2.2 shows the properties of each recorded sequence.

All sequences were recorded in the same room, using two GoPro cameras (HERO3 White edition). The fisheye effect was corrected using the GoPro Studio software tool. Each camera views are shown in Figure 2.5 and a room top view map is shown in Figure 2.6.

The ground truth of this dataset was manually annotated for each frame of each sequence. The annotated ground truth considers the wheelchair users and the standing people present in every frame, even if they are highly occluded. This dataset and its annotated ground truth

are publicly available for research purposes in the Wheelchair users dataset webpage (http://www-vpu.eps.uam.es/DS/WUds/).

Fig. 2.5. Camera views of the Wheelchair Users dataset. Left: viewpoint 1. Right: viewpoint 2. Up: Empty room. Down: examples with people.

Sequence number	#Frames	#Wheelchair users	#Standing people
1	1318	1	0
2	916	1	0
3	860	1	1
4	1167	1	1
5	1638	2	0
6	723	2	0
7	1082	2	2
8	743	2	2
9	2102	2	2
10	2460	2	2
11	1855	4	0

Table 2.2: Properties of each of the recorded sequences from the Wheelchair Users dataset.

In addition, in order to evaluate not only the yes/no detection decision but also the precise people locations, we take into account the three evaluation criteria defined in [Leibe et al., 2005], that allow to compare hypotheses at different scales: relative distance (dr), cover and overlap. A detection is considered true if $dr \leq 0.5$ (corresponding to a deviation up to 25% of the true object size) and cover and overlap are both above 50%.

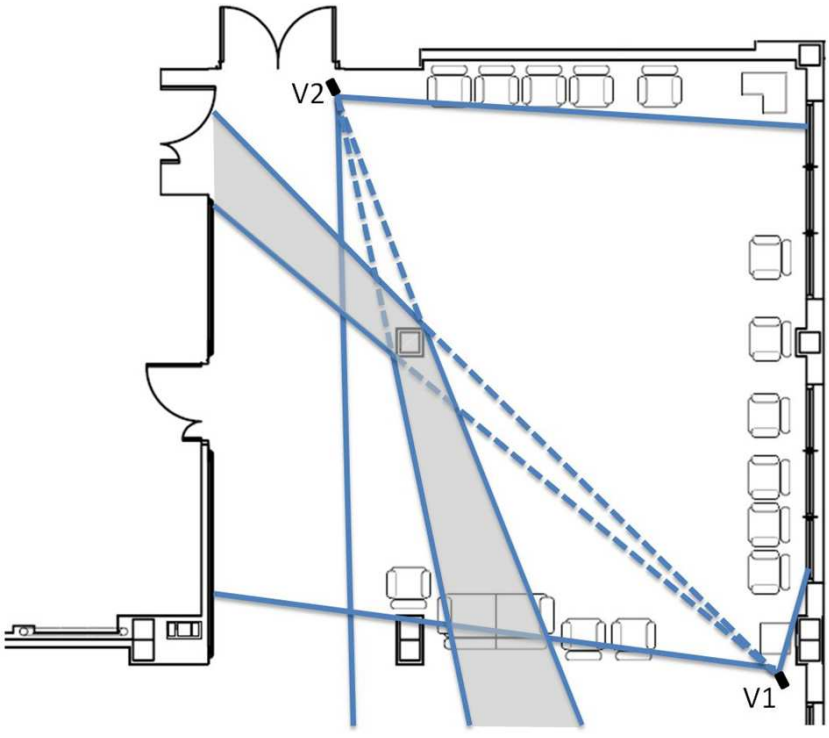


Fig. 2.6. Top view map of the Wheelchair Users dataset. V1 and V2 represent camera 1 and camera 2 locations and the fields of view of each camera are represented.

Table 2.3: Properties of each of the image sets from the parking lot dataset.

Sequence name		#Frames	#Vehicles
Training		6616	28231
Test	All_Cam1	1000	12275
	All_Cam2	1000	9738
	Synchronized_Cam1	100	751
	Synchronized_Cam2	100	749

2.3.2 Parking Lot dataset

The Parking Lot dataset (PLds) was recorded as there was a lack of public parking lot datasets. The sequences were recorded in a real environment (Pittsburgh International Airport parking lot), in order to work with an environment as realistic as possible. Each frame, recorded using *Panasonic WV-SW155* cameras, has a resolution of 1280×960 pixels. Figure 2.7 shows an example of each one of the two viewpoints (2.7a and 2.7b), and examples of different illumination (day, night, sunrise with shadows) and weather (sunny, rainy) conditions.

The dataset consists of two main image sets: a training set which consists of a longer set of images (6616 frames) and the test set which consists of a long (named All_CamX) and a short (named Synchronized_CamX) version of the images with 1000 and 100 frames, respectively. The short versions (Synchronized sets) are subsets of the long versions: they consist of frames synchronized between the two cameras, to be able to evaluate the multi-camera setup. The different image sets details are presented in table 2.3. The synchronization of the images has been performed by Optical Character Recognition (OCR) applied to the date and time recorded in each camera, and selecting frames with the greatest possible variability of climate and lighting conditions.

In addition to generating the images, the vehicles of all images have been manually annotated. The training images have been annotated for its use in the generation of the parked vehicle model, and the test images for the evaluation of the parking vacant management system. In the case of the Synchronized set, the vehicle occupancy matrix has been manually generated to allowfor evaluating the system at this level.

This dataset and its annotated ground truth are publicly available (http://www-vpu.eps.uam.es/DS/PLds/).

2.4 Generation of synthetic datasets

Apart from describing the existing and proposed datasets, we propose a new technique for the generation of synthetic datasets to treat the problem of generating an object detection model with different appearance of an existing semantic object class model. The main idea of this

Fig. 2.7. Examples of dataset frames: (a) shows an example of Camera 1 viewpoint, (b) shows an example of Camera 2 viewpoint. (c)-(f) show examples of different illumination and weather conditions.

technique is to generate images of an object/person with a different appearance by combining patches (with different edge processing or segmentation) and its appropriate rescaling. The set of generated images will be used later to train a detection model.

Due to the extension and details given for this technique, and due to it has been evaluated to analyze a proposed and developed example, this technique is presented in greater detail in Chapter 3. The motivation of that chapter is the study of feasibility of training an object detection algorithm using a synthetic images dataset.

2.5 Existing metrics

2.5.1 Precision, Recall, PR-Curve and F-score

In order to evaluate the different detection approaches, we quantify the performance results. Global sequence performance is usually measured in terms of Precision-Recall (PR) curves [Leibe et al., 2008; Andriluka et al., 2008; Wojek et al., 2009], which is a metric widely used in pattern recognition to validate results. These curves compare the similarities between the output and ground truth bounding boxes. In our case, we consider the output bounding boxes of the detectors. For each value of the detection confidence or score, Precision-Recall curves are computed:

$$\text{Precision} = \frac{\#TPD}{\#TPD + \#FPD} \tag{2.1}$$

$$\text{Recall} = \frac{\#TPD}{\#TPD + \#FND} \tag{2.2}$$

Where TPD are True Positive Detections, FPD are False Positive Detections, and FND are False Negative Detections.

In addition, in order to evaluate not only the yes/no detection decision but also the precise people locations, we take into account the three evaluation criteria defined in [Leibe et al., 2005], that allow to compare hypotheses at different scales: relative distance (dr), cover and overlap. A detection is considered true if $dr \leq 0.5$ (corresponding to a deviation up to 25% of the true object size) and cover and overlap are both above 50%.

The integrated Average Precision (AP) is generally used to summarize the algorithm performance in a single value, represented geometrically as the area under the PR curve (AUC-PR). In order to approximate the area correctly, we use the approximation described by [Davis and Goadrich, 2006]. The greater the area under the curve, the better the performance.

F-score considers both the precision and the recall. The F-score is the harmonic average of the precision and recall, reaching its best value at 1 (perfect precision and recall) and worst at 0. It is calculated as:

$$\text{F-score} = \frac{2 \cdot \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3}$$

2.6 Proposed and adapted metrics

2.6.1 Ground plane evaluation

In a complementary way to the previous evaluation, a ground plane level evaluation has been considered in order to evaluate object detections in an additional dimension (depth), considering the floor plane instead of just each camera viewpoint. For this evaluation, all ground truth detections are projected on the ground plane, thus obtaining the ground truth of this evaluation. Each projected blob is defined as a point (center) and a radius (the same operation as that applied when obtaining the cylinders in Chapter 6). The detections of each camera are also projected on the ground, defined in the same way as the bounding boxes of the ground truth. An example of this circumferences is shown in Figure 2.8.

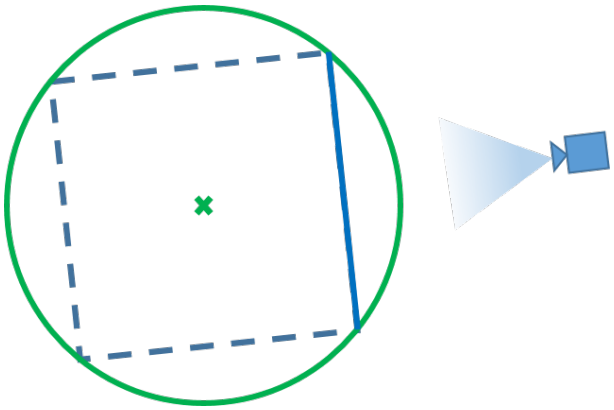


Fig. 2.8. Circumference diagram for ground plane evaluation. The continuous blue line represents the projection of the bounding box.

For the evaluation, ground truth and detections bounding boxes are associated using the minimum distance between pairs, if, and only if, there is any spatial overlap between the cylinders associated with each bounding box. Through this association, the true positives, the false negatives and the false positives are defined, allowing for the calculation of Precision and Recall scores, using the same formulas presented in the previous subsection. Finally, the PR curves are obtained, extracting the AUC value from them. Figure 2.9 presents an example of the described ground plane evaluation, showing examples of the described associations.

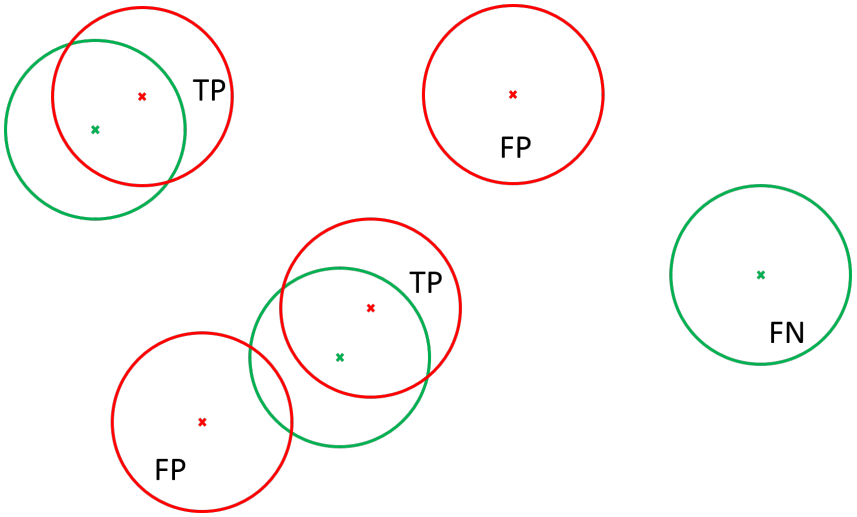


Fig. 2.9. Example of ground plane evaluation. Green circles represent ground truth. Red circles represent detected objects. TP, FP and FN represents True Positives, False Positives and False Negatives, respectively.

2.6.2 Parking spots evaluation

Global sequence performance is usually measured in terms of Precision-Recall (PR) curves [Andriluka et al., 2008; Leibe et al., 2008; Wojek et al., 2009]. These curves compare the similarities between the output and ground truth bounding boxes.

For the parking spots evaluation, two uses of the evaluation metrics are distinguished. The first is the common one used for object detection, previously described in Subsection 2.5.1. The second use is at occupied/empty spots level, according to the occupation matrix of the parking lot. Parking spaces may be occupied or empty. In this case, it is considered a classification for each place, and the overlap is not measured for it. The occupation matrix and the ground truth are compared to define true positives, false positives, false negatives and true negatives, as shown in Table 2.4.

Table 2.4: Occupation matrix evaluation table.

Detected spot status	Ground truth status	Spot evaluation
Vacant	Vacant	True Negative
Vacant	Occupied	False Negative
Occupied	Vacant	False Positive
Occupied	Occupied	True Positive

2.7 Existing detection algorithms

Unlike datasets and metrics, in this chapter only existing algorithms are presented which will be later considered to apply the proposed techniques described in the rest of the chapters. The presented algorithms are: DPM, ACF, Faster-RCNN and YOLO9000.

The first considered detection algorithm is the Deformable Parts Model (DPM) detector [Felzenszwalb et al., 2010b]. The DPM detector is based on exhaustive search and a part-based person model. It is a part-based adaptation of the original Histogram of Oriented Gradients detector (HOG) [Dalal and Triggs, 2005]. It proposes an object detection system based on mixtures of multiscale deformable part models where each deformable body part is modeled as the original HOG detector [Dalal and Triggs, 2005]. The algorithm model also contains the flip (horizontally mirrored) of the model.

The second considered detection algorithm is the Aggregated Channel Features (ACF) detector [Dollar et al., 2014]. Multi-resolution image features are approximated via extrapolation from nearby scales allowing to design an object detection algorithms that is as accurate as previous approaches, and considerably faster. It compute finely sampled feature pyramids at a fraction of the computational cost, without sacrificing performance: for a broad family of features this approach find that features computed at octave-spaced scale intervals are sufficient to approximate features on a finely-sampled pyramid.

The third considered detection algorithm is the Faster-RCNN (Regions with Convolutional Neural Network Features) [Ren et al., 2015] detector, which consist in a more efficient variation, mainly in terms of computational cost but also in performance, of the previous versions R-CNN [Girshick et al., 2013] and Fast R-CNN [Girshick, 2015] detectors. The three variations have in common the combination of bottom-up region proposals with rich features computed by a convolutional neural network. The main difference of the Faster-RCNN is the use of a Region Proposal Network (RPN) that enables nearly cost-free region proposals.

The last considered detection algorithm is the YOLO9000 [Redmon and Farhadi, 2017], a real-time object detection system that can detect over 9000 object categories, which improves the previously published YOLO detection method [Redmon et al., 2016]. This approach considers object detection as a regression problem with spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. As the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance. The improvement of YOLO9000 over the previous version is that this one focus on improving recall and localization while maintaining classification accuracy, as YOLO makes a significant number of localization errors.

2.7.1 Detection algorithms considered for each chapter

This subsection lists the chapters in which each detector has been considered.

First, DPM detector has been considered for Chapters 3, 4, 5, 6 and 7. This is the first object detection algorithm that was known and therefore has been applied in all of them. ACF detector was subsequently considered. Tests were carried out on the system presented in Chapter 5, but it was discarded due to detections covered the roof of the vehicle instead of the complete vehicle and for that reason it did not fulfill one of the main needs of the detections considered for the system. However, this detector was later considered for Chapter 7. Faster-RCNN detector is evaluated for Chapters 4, 5 and 7. Finally, YOLO9000 was recently published and has only been applied to the last developed contribution, presented in Chapter 7.

The cases of not using algorithms for chapters that could consider them, which have not been explained their non-inclusion, are mainly due to shortage of time and deadlines.

2.8 Conclusions

This chapter has presented an evaluation framework that allows rigorously and objectively evaluate the systems, algorithms or techniques developed in the following chapters.

With respect to datasets, existing datasets that met our analysis of requirements have been selected, which have been completed with the creation of two new datasets, completing a total of five datasets considered for the evaluation of, on the one hand, people with different appearances (standing or in wheelchair) and, on the other hand, objects (especially people, but also vehicles or any other object that desired to be detected) considering information from multiple cameras. In addition, the idea of generating synthetic images has been introduced, which will be developed in the next chapter.

With respect to evaluation metrics, this chapter has presented metrics commonly used for the evaluation of object detectors. First, the "classical" evaluation metrics have been formulated, and then we have adapted these metrics for, on the one handconsidering a third dimension (depth) in the scenarios, and on the other hand evaluating the capacity to detect occupied or empty parking spots.

Finally, with respect to detection algorithms, four algorithms have been presented, which use different methods and features for object detection. In the following chapters these algorithms are used, so their results can be compared.

Chapter 3

Generation and evaluation of synthetic models for training people detectors

3.1 Introduction[1]

There is a large demand in the area of video-surveillance, especially in people detection, which has caused a large increase in the amount of researches and resources in this field. As training images and annotations are not always available, it is important to consider the cost involved in creating the detector models. For example, for elderly people detection, the detector must take into account different positions such as standing, sitting, in a wheelchair, etc. Therefore, this work has the main objective of reducing the amount of resources needed to generate the detection model, saving the cost of having to record new sequences and generate the associated annotations for a detector training.

The performance of people detectors varies depending on the environment in which they are tested, as they have a great dependence on factors such as illumination, occlusions, person pose, perspective, distance from the camera, etc. The motivation of this chapter is the study of feasibility of training an object detection algorithm using a synthetic images dataset. In particular, the chosen object to train is a wheelchair user from empty wheelchairs images and standing people images. Three synthetic image datasets have been created in order to train three different models, evaluating which model is optimal and finally analyzing its feasibility by comparing it with a people detector for wheelchair users trained with real images. Other people detection scenarios in which this technique could be applied are, for example, people riding horses or motorbikes, or people carrying supermarket carts. The synthetic datasets have been

[1]This chapter is an adapted version of the publications [R. Martín-Nieto and Martínez, 2017]

generated by combining images of standing people with wheelchair images, combining image patches, and segmenting sections of people (trunk, legs, etc.) to add them to the wheelchair image. As expected, the obtained results have a reduction in accuracy (between 21 and 25%) in exchange for the enormous saving in human annotation and resources to record real images.

The structure of this chapter is as follows: after this introduction in Section 3.1, people detection related works are presented in Section 3.2. Section 3.3 describes the synthetic dataset creation methods, and the experiments and results are shown in Section 3.4. Finally, Section 3.5 presents the conclusions and the future work.

3.2 People detection: related work

People detection has become one of the research areas of greatest interest in the field of image and video processing. Several different detection systems have been developed, however, as explained in [García-Martín and Martínez, 2015b], most people detectors have a common architecture, which consists of the design and training of a person model, based on certain parameters, such as movement, silhouette or posture. The next step is to adapt this model to all possible candidates to be a person in the scene, and finally, if that candidate fits the model, it will be classified as a person, while those that do not fit will not be classified as person.

3.2.1 Systems architecture for people detection

The main stages of the architecture of a generic people detector are described below:

- Input: there are many possible formats, however, for computer vision, the basic input unit are images or frames.
- Object detection: consists of the generation or extraction of the possible initial candidates (locations) to be a person. It is a critical task for the detector. It will be explained in Subsection 3.2.2.
- Person model: it defines the features and rules that the objects must fulfill to be considered as a person.
- Verification or Classification: it has the same operation as a pattern detector, comparing previously trained models with the model generated from the sequence.
- Decision: using the result of the previous stage, this one decides if the detected object is (or is not) a person.

3.2.2 Object detection

There are two main approaches to object detection [García-Martín and Martínez, 2015b]: the first one, segmentation, focuses on foreground and background information, and the second one, which is based on exhaustive exploration. Both approaches will be explained in more detail below. In spite of having different approaches, the final result for both is the location and the dimension of the different objects detected in the scenario.

- Segmentation: used to divide the image into different regions, which ideally correspond to different objects in the real world. This process tries to assign a label to all pixels, so that pixels with the same label share some visual characteristic, such as color, movement, texture, etc. Contiguous regions must have very significant differences with respect to the same feature to be considered a different region. This technique tries to locate and discriminate objects from the foreground with respect to the background, as done in García-Martín et al. [2012].

- Exhaustive search: consists of a sweep of the image to find similarities with the chosen person model, at different scales and in different positions. A very dense confidence map is obtained with this approach, therefore to reach individual detections a search for local maxima in the density volume should be performed, and then some kind of non-maxima suppression should be applied. There are two techniques for this, the first one obtains this volume of density by evaluating different detection windows with a classifier, as is the case of detectors based on sliding window, i.e. [Alonso et al., 2007], while the second one generates this density volume by probabilistic votes issued by equivalent local features. This technique is the one used by feature-based detectors, such as [Leibe et al., 2007, 2008].

- Segmentation and exhaustive search: combines the two previous techniques, trying to take advantage of their respective strengths. Initial candidates are selected with the segmentation method in the first round and then performs a second round through exhaustive search.

3.2.3 Standing people detection

In computer vision, standing people detection can be considered as a two steps process [Hu et al., 2004; Valera and Velastin, 2005]. First, it is necessary to localize the initial objects candidates to be standing people in the scene. The two most common approaches to localize those objects are those based on some kind of segmentation of the scene in foreground (objects) and background [Kilambi et al., 2008] and those based on a scanning approach [Enzweiler and Gavrila, 2009; Felzenszwalb et al., 2010b]. In general, those algorithms based on a scanning approach have been

proved to be more robust to real and more complex sequences where there are several background and people variabilities [Enzweiler and Gavrila, 2009; Dollár et al., 2012b; Gerónimo et al., 2010; Simonnet et al., 2012]. There are also some approaches that try to combine both approaches together [Alonso et al., 2007; García-Martín and Martínez, 2010].

The second step in any standing people detection can be considered as a standard pattern recognition issue. In this case, it is necessary to previously define a standing person model and then classify any new candidate selected during the previous step as a standing person or not. The classification process will be characterized according to the chosen standing person model. Therefore, standing people detection approaches can be classified into two groups, namely, holistic and part-based detectors, depending on the model properties. The holistic detectors define the person as a region or shape [Dalal and Triggs, 2005; Dollár et al., 2012a; Leibe et al., 2008; Viola and Jones, 2004], whilst the part based detectors define the person as combination of multiple regions or shapes [Felzenszwalb et al., 2010b; Andriluka et al., 2009]. In general, those algorithms based on part-based models are able to deal with partial occlusions better than those based on a holistic model, but significantly increasing the model complexity.

In recent years the object detection results (and therefore people detection results) have been greatly improved thanks to the use of deep learning algorithms. Some examples of these algorithm are [Ouyang et al., 2014], [Girshick et al., 2014] or [Ren et al., 2015].

3.2.4 Wheelchair users detection

There are some works in the state of the art trying to address the wheelchair users detection problem. These works can be classified into two main groups. The first group focuses on detecting ellipses which correspond to the wheelchair wheels. The second group is based on detecting the wheelchair users using discriminative features, usually color and Histogram of Oriented Gradients (HOG).

The first approach of the works that try to find the wheel ellipses is presented in [Myles et al., 2002]. The model considered here is based on two wheels with a head over them. The wheels are detected using the Hough transform to detect ellipses in an edge image obtained via the Canny detector. The head is found using a skin detector. All these stages are performed after a background subtraction. In [Yang and Chung, 2007], the detection is based only in determining the location and orientation of the wheels, proposing a mathematical method of ellipse-circle geometry. [Wu et al., 2010] follows the work presented in [Myles et al., 2002] and includes tracking and event detection. In this case, Zimmer frames are also detected. The location of doors is also used for the detections. The wheelchair users detector presented in [Huang et al., 2013b] starts from a background subtraction stage, similar to [Myles et al., 2002]. After obtaining the foreground, the resulting bounding boxes are analyzed locating the wheel, and then the user and the assistant (if any). A novel idea is presented in this work, which is to

recognize whether an assistant is pushing the wheelchair.

On the other hand, the second group aims to find discriminative features to detect the wheelchair user. Similar to other studies, [de Chaumont et al., 2004] starts with a background subtraction. This solution is based on detecting wheelchair user parts (e.g., head, chest, legs) and wheelchair parts. Finding each part is based on color. After that, the object position is obtained using a stereo vision camera. The justification for not trying to locate the wheels is that there are orientations (front and rear) in which they are unobservable. Besides, there are different wheelchair models, especially electrical, which do not have large wheels like conventional wheelchairs. The recognition proposed in [Hosotani et al., 2009] also uses stereo vision cameras. The feature used is HOG, allowing discrimination between standing people and wheelchair users thanks to a previously trained Support Vector Machine (SVM). The detector proposed in [Huang et al., 2010] considers two descriptors, HOG and Contrast Context Histogram (CCH), which are adopted to model, respectively, the shape and appearance of the wheelchair. An AdaBoost learning stage selects the features which better discriminate the object. All the possible wheelchair orientations are classified in 8 different models composing a state graph whose elements can change to adjacent orientation models. A Gaussian pyramid is constructed to overcome the scale problem by downsampling the image from the original resolution. The approach proposed by [Huang and yu Chen, 2012] focuses on a dimensionality reduction using sparse representation to improve the generalization capability. To characterize the wheelchair users, directional maps are defined by determining the dominant direction of motion in each local spatiotemporal region.

3.3 Synthetic dataset creation methods

It is interesting to study the feasibility of creating a detector with synthetic images. In this section, the standing people images will be combined with wheelchair images, using different methods that will be detailed later. Obtaining a reliable detector in this way would avoid the cost of having to record a large number of images, since, combining them would result:

$$\#Images_{wheelchairuser} = \#Images_{sp} \times \#Images_{wh} \tag{3.1}$$

Where $\#Images_{wheelchairuser}$ is the resulting number of wheelchair user images, $\#Images_{sp}$ is the number of standing people images, and $\#Images_{wh}$ is the number of wheelchair images.

In our study case, 3600 images were necessary to train the original (no synthetic) wheelchair user model (see Chapter 4). In order to obtain comparable detection results, it was decided to achieve a similar number of images to the one used to generate the original wheelchair user model. To get closer to this number, 75 people images and 45 wheelchair images were obtained, resulting in a total of 3375 images, which is a close number to the one used for the original model. The images selection was made with the intention of fulfilling certain characteristics to

(a) (b) (c)

Fig. 3.1. Image examples for each generated dataset: (a) Basic combination, (b) Edge smoothing combination and (c) Masked combination.

get resulting images as close as possible to a real image. These characteristics are:

- The selected person could not be occluded.
- The legs should not be separated.
- The person should be facing the camera or sideways, never on his back.

Apart from this selection, 45 wheelchair photos from different sources were collected. Once obtained, patches were selected from both the torso and the legs, from the standing people images, and were combined with the wheelchairs images using three different methods, thus creating three datasets. These methods are described below:

1. Basic combination: both the torso and the legs of the person are selected and are placed in the wheelchair image where they should intuitively be. Special care was taken to ensure that images to be combined were as realistic as possible. The position of these parts was annotated in each wheelchair image as well as in each person image, although in this case only the torso and the lower part of the legs were annotated. Apart from this annotation, measurements of both the hip and the distance from one knee to another, for both sets (wheelchair and person), and from these annotations the patch is calculated rescaling the annotated patch to fit the width that must have in the image. An example of the image obtained for this dataset is shown in Figure 3.1(a).

2. Edge smoothing combination: in Figure 3.1(a) it can be seen that due to the image combination, the edges of the body and legs patches are very significant, which should be avoided since it does not correspond to the real object. Trying to diminish the effect that these marked edges produce on the model, it was decided to smooth the edges. This smoothing was performed by a smoothing filter with a 9x9 pixels kernel. The result of this combination can be seen in Figure 3.1(b).

3. Masked combination: to give more realism to the set of images, we tried to eliminate the background corresponding to the image of the standing person, fitting and masking the patches to the person part silhouette. The final image is the result of combining the wheelchair image with the person body (trunk and head) and legs, as can be seen in Figure 3.1(c). Masks were used to eliminate the background information. An example of such employed masks is shown in Figure 3.2. Those masks corresponds to the silhouette of the person, having the inside of the silhouette value 1, and the outside value 0. Taking advantage of these masks, each image patch was combined as follows:

$$I_{final} = I_{wheelchair} * (1\text{-}I_{mask}) + I_{person} * I_{mask} \tag{3.2}$$

(a) (b)

Fig. 3.2. Standing people image example (a), and standing people mask example (b). Extracted from http://www-vpu.eps.uam.es/DS/CVSG/.

After completing the three datasets, the models were created using the Deformable Parts Model (DPM) [Felzenszwalb et al., 2010b] detector training code, resulting in Figures 3.3(a), (b) and (c). Observing the three final models, similarities between them can be found. In spite of the different techniques, the final result is very similar, although they have small differences. The head region is better appreciated with the masked combination technique. More border artifacts appear in the basic combination model than in the edge smoothing combination model. Another difference between the combination with mask application and the other two, is the greater degree of detail that is seen in the chair. It is interesting to see how the lower edge of the upper body, and the upper edge of the legs are detected as edges and included in the model. In addition, the similarity between the model trained with real images (Figure 3.3(d)), and the models trained with the synthetic dataset (Figures 3.3(a), (b) and (c)), can be observed..

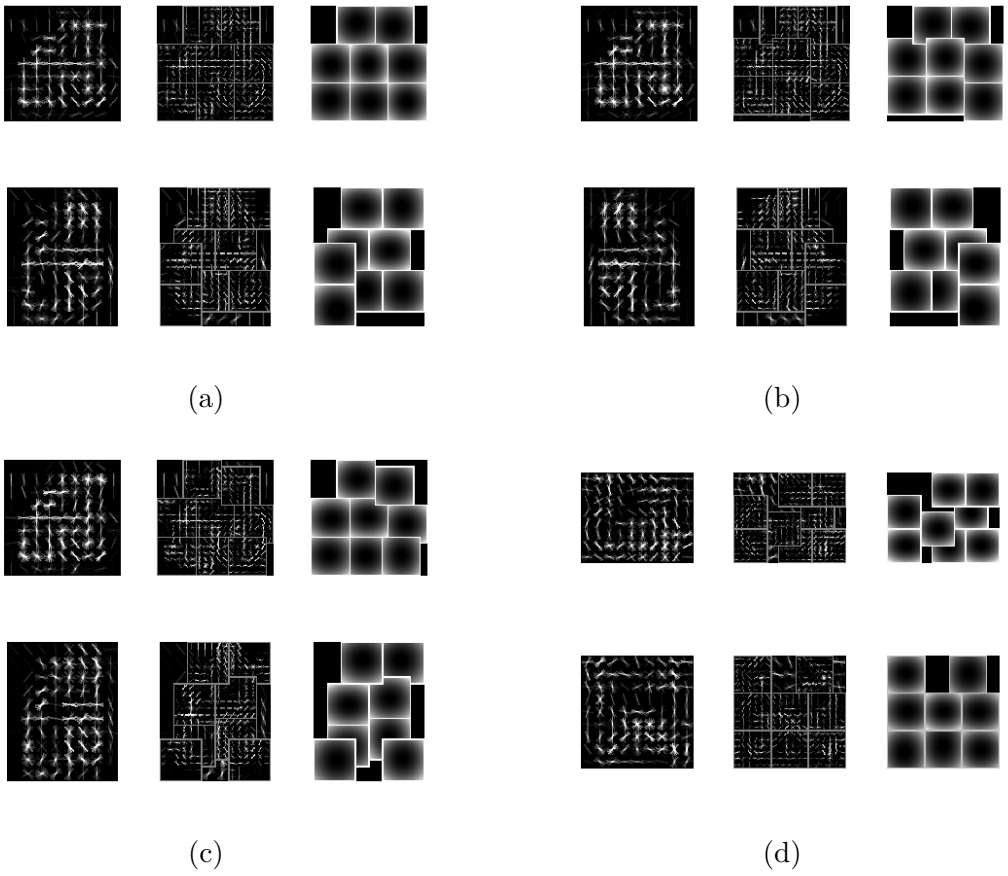


Fig. 3.3. Generated models: (a) basic combination, (b) edge smoothing combination, (c) masked combination and (d) real images model.

Table 3.1: AUC values for the three generated detector models.

	Sequence1	Sequence2	Average	Percentage change
Basic comb.	0.649	0.562	0.605	-24.48%
Smooth comb.	0.691	0.586	0.639	-21.15%
Masking comb.	0.588	0.595	0.592	-25.86%
Real images	0.93.2	0.768	0.850	

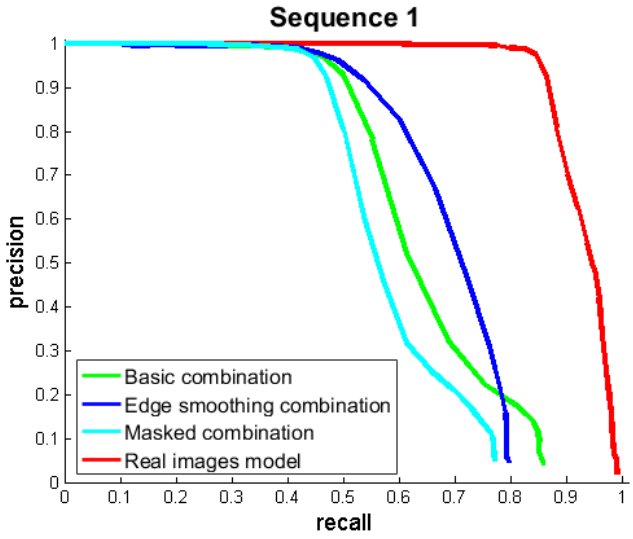
3.4 Experiments and results

In this section the obtained results by the different trained detector models are presented: basic combination, edge smoothing combination and masked combination. The Wheelchair Users dataset (see Subsection 2.3.1) was used to evaluate the generated models. As can be seen in the precision-recall curves shown in Figure 3.4(a), for sequence 1 the detector with the highest AUC is the detector which is trained with the edge smoothing combination images, and the one with the lowest AUC is the detector based on masked combination. On the other hand, in the case of Figure 3.4(b) the curve with the best AUC is the one which corresponds to the edge smoothing combination model, and the worst AUC score is obtained by the basic combination model. On average, the AUC scores are very close for the three trained models, being slightly better the model based on edge smoothing.

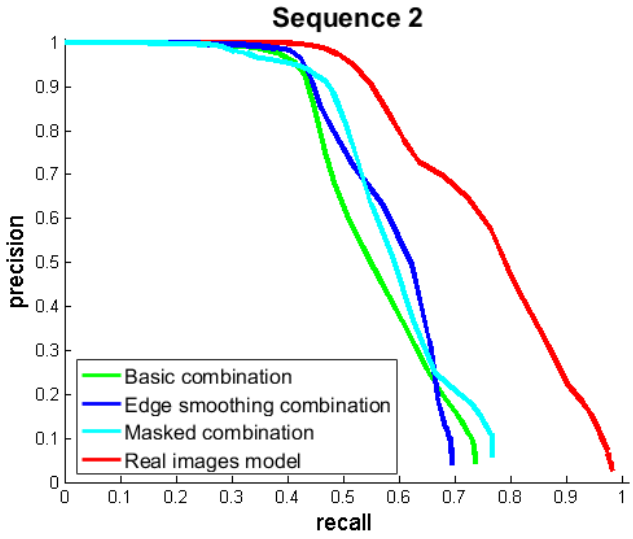
This result was expected because smoothing the edges of the patch reduces the impact of artifacts in those areas where the detection model is generated. As the object detection algorithm is based on edge search (HOG), this benefits the performance of the model. The masked combination has probably not improved due to as it fits so much the person it adds edges that should not be added to the model. In spite of this, this model obtains points of the curve better than those obtained by the other two models for certain thresholds in sequence 2.

3.5 Conclusions

Three synthetic image datasets have been created in order to train three different models, evaluating which model is optimal and finally analyzing its feasibility by comparing it with a people detector for wheelchair users trained with real images. The performance of the trained models for the two sequences is not exactly the same, although on average the performance is similar. Looking at the average performance, the best result is obtained with the edge smoothing combination, while the worst is obtained with the mask combination model. With these results it can be concluded that the performance of the detectors is acceptable, although worse than the one obtained with the original wheelchair people detector, trained with real images. This result was expected a priori, as the images that have been generated are synthetic and are different from the real recorded images, but despite this, a detector model has been obtained that is able



(a)



(b)

Fig. 3.4. Precision-recall curves for the generated detector models: basic combination (green), edge smoothing combination (blue), masked combination (cyan) and real images model (red).

to detect in an appropriate way, with results of areas under the curve between the 50% and 60% of the AUC.

In exchange for this performance loss, a functional detector has been obtained without the need to record the real object (in this case, wheelchair users). This method could be useful in situations where it is not possible to compile or record a dataset of the desired object type, or obtain it is too expensive in terms of time or resources.

These first approach results are promising and can be improved by generating other more elaborated synthetic image datasets. Observing the masking combination model, some edges are obtained in the area where the wheelchair and the legs join. Smoothing that edges can improve the model. Adding a waist patch can give more realism to the resulting image which can result in a better model. It would be also interesting to test other combinations such as the ones mentioned previously in this chapter: people riding horses, people with shopping carts, etc. Finally, a detection model generated from real images could be completed with this set of generated images in order to improve its detection capacity.

Part III

Detection Approaches and Applications

Chapter 4

Incorporating wheelchair users in people detection

4.1 Introduction[1]

In health care centers, senior residences, hospitals, etc., it is usual to see people who need wheelchairs and their detection is useful to monitor them and to provide them assistance in case they need it. Knowing the location of a wheelchair user can be useful for some healthcare applications (e.g. monitoring) and it can be used to analyze the behavior and actions of such users in different environments. The automatic detection of mobility impaired people, including wheelchair users, is also an important problem for Intelligent Transportation Systems (ITS) in public traffic areas [Hosotani et al., 2009]. Many assistance applications can be derived from automatic wheelchair users detection, e.g., doors, elevators, escalators, can automatically activate a special operation mode for such people after detecting them, or the green-light time can be increased in pedestrian crossing with traffic lights when a wheelchair user is detected. All these events could be activated manually by one person, but, if automatic activation is achieved, people in wheelchairs would feel more comfortable and these events would become something natural, and the operation would not need human agents for correct functioning.

An application environment for which the presented detector is useful is independent living. According to the definition given by the World Institute on Disability (http://www.wid.org/), independent living is defined as allowing people with disabilities to have the same level of choice, control and freedom in their daily lives as anyone else. In the context of caring for the elderly, independent living is seen as a continuum care, whose next step would be the incorporation to a nursing home. The proposed wheelchair user detector is useful for both stages, first to monitor the wheelchair user in their domestic environment ensuring that everything runs properly, and

[1] This chapter is an adapted version of the publications [Martín-Nieto et al., 2018a]

then to video monitor people in a nursing home, allowing to detect interesting events such as fall detections [Auvinet et al., 2011; Bian et al., 2015].

A wheelchair users detector is presented to extend people detection, providing a more general solution to detect people in environments such as houses adapted for independent and assisted living, hospitals, healthcare centers and senior residences. A wheelchair user model is incorporated in a detector whose detections are afterwards combined with the ones obtained using traditional people detectors (we define these as standing people detectors). We have trained a model for classical (DPM, [Felzenszwalb et al., 2010b]) and for modern (Faster-RCNN, [Ren et al., 2015]) detection algorithms, to compare their performance. A final application is shown on which the detectors output is combined generating a trajectory for each standing and sitting person, projecting it on the plane of a nursing home.

The structure of this chapter is as follows: after this introduction, Section 4.2 presents existing works and publications related with the people detection, including wheelchair users detection. Section 4.3 describes the detection approach. The evaluation and results of the approach are presented in Section 4.4 and an example of application is shown in Section 4.5. Finally, Section 4.6 contains conclusions and future work.

4.2 State of the art

For this section, the related works are the same as those presented in Subsection 3.2.3 for the standing people detection, and Subsection 3.2.4 for the wheelchair users detection.

The proposed wheelchair users detectors (see section 4.3.1) have advantages over the previously existing solutions: it does not need a background model for background subtraction, it can detect wheelchairs in any orientation, it does not need to know the dimension of some parts of the wheelchair, it does not need stereo vision cameras, it does not need to know the wheelchair colors in advance, and it does not consider the wheelchair user as a rigid object, allowing deformations.

We have chosen a scanning approach with a part-based model (DPM, [Felzenszwalb et al., 2010b]), and a deep learning approach (Faster-RCNN, [Ren et al., 2015]) for the object detection algorithms. We have chosen these two detection algorithms as the first one, DPM, is a classic algorithm, based on HOG filters, that offers good detection after the great improvement of the detection algorithms in the last years, and the second one, Faster-RCNN, to observe the operation of the proposed technique using one of the most modern and effective algorithms of the state of the art, based on neural networks.

4.3 Detection approach

This section describes the original detection algorithms (see section 4.3.1), whose training method is used to generate the wheelchair users detection models (see section 4.3.2). The detections from the different models (standing people and wheelchair users) are combined for an integrated detection (see section 4.3.3).

4.3.1 Detection algorithms

The first considered detection algorithm is the Deformable Parts Model (DPM) detector [Felzenszwalb et al., 2010b]. The DPM detector is based on exhaustive search and a part-based person model. It is a part-based adaptation of the original Histogram of Oriented Gradients detector (HOG) [Dalal and Triggs, 2005]. It proposes an object detection system based on mixtures of multiscale deformable part models where each deformable body part is modeled as the original HOG detector [Dalal and Triggs, 2005]. The algorithm model also contains the flip (horizontally mirrored) of the model.

The second considered detection algorithm is the Faster RCNN (Regions with Convolutional Neural Network Features) [Ren et al., 2015] detector, which consist in a more efficient variation, mainly in terms of computational cost but also in performance, of the previous versions R-CNN [Girshick et al., 2013] and Fast R-CNN [Girshick, 2015] detectors. The three variations have in common the combination of bottom-up region proposals with rich features computed by a convolutional neural network. The main difference of the Faster-RCNN is the use of a Region Proposal Network (RPN) that enables nearly cost-free region proposals.

The computational cost of the detections is not treated in this chapter as this aspect is analyzed by the authors of DPM ([Felzenszwalb et al., 2010b]) and Faster-RCNN (Ren et al. [2015]). The used DPM approach is implemented with MATLAB and the computational cost is about 2 seconds per frame, considering an image of 352 $\times$ 288 pixels. Note that there is also a faster implementation in OpenCV that improves the detection time to about 1 second per frame. The used Faster RCNN approach is implemented with MATLAB and Caffe [Jia et al., 2014], and the computational cost is about 150-200 milliseconds per frame (Faster RCNN, VGG-16 with GPU), considering an image of 500x375 pixels.

4.3.2 Detection models

This subsection adds some details about the two different trained algorithm models. The standing people model has not been trained for this work, but it is presented here for comparing it with the wheelchair users model.

Figure 4.1 shows a visual example of the DPM person model, namely the INRIA person model, extracted from [Felzenszwalb et al., 2010a]. The model also contains the flip of the

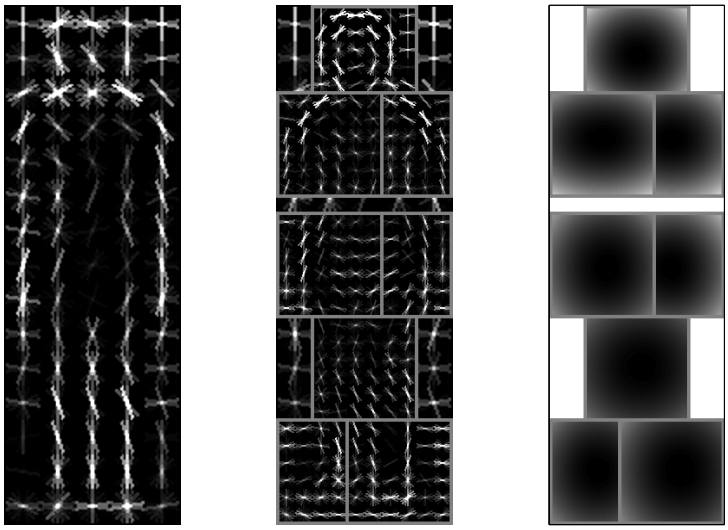


Fig. 4.1. DPM standing people model. The three columns are, from left to right, root model, parts model and parts deformation.

model, but it has not been included in the figure as it does not provide additional information different from the data already shown.

Following the original people detection algorithm, we train a wheelchair users detector model.

To generate the wheelchair users model, we used the annotations of the training set from the Smile Lab training dataset (see subsection 2.2.3), containing 3674 positive examples. For the negative examples set, we used the standing people model negative examples from [Felzenszwalb et al., 2010a]. For this purpose, we ensure that this image set does not contain any pictures with a wheelchair nor a wheelchair user.

For the DPM standing people detector model, there is just one model variation as the appearance from the different points of view are similar. Unlike the standing people model, a model with two variations is trained for the wheelchair users, as it is considered that the appearance of the front and side wheelchair users are different enough to be independent in their appearance classification. We have also performed experiments testing from 2 to 8 model variations, obtaining very similar or worse results, due to the overfitting of the model to the training data. Figure 4.2 shows the resulting wheelchair user model. The trained model also contains the flip of each model (as the original people detector model), but it has not been included in the figure as again it does not provide additional information different from the data already shown.

For the Faster-RCNN detector, and according to the author's results [Ren et al., 2015], we

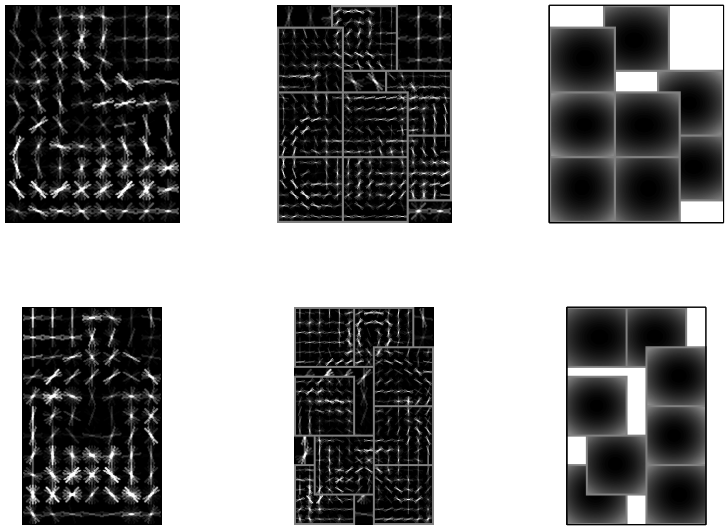


Fig. 4.2. DPM wheelchair user model. Each row represents a model variation. The three columns are, from left to right, root model, parts model and parts deformation.

have chosen the pre-trained network VGG-16 model [Simonyan and Zisserman, 2014] that has 13 convolutional layers and 3 fully-connected layers. We have refined the network weights using the PASCAL VOC 2007 and 2012 datasets, and we have added a new object class, the wheelchair user object, using the same positive and negative examples than for the DPM model training, from the SmileLab wheelchair dataset (see Subsection 2.2.3). The Faster-RCNN model does not have a graphic representation as in the case of the DPM model.

This wheelchair users models are available for research purposes in the Wheelchair users dataset webpage (`http://www-vpu.eps.uam.es/DS/WUds/`).

4.3.3 Detectors combination

The DPM wheelchair user detections and the standing people detections are combined to obtain the general people detections. All the detections from each detector are maintained as we consider that each detector works for disjoint people models. As each detector has a different Standing People (SP) / Wheelchair User (WU) Detection Confidence output space or range $C_{SP/WU}$ (see Figure 4.4), in order to add the outputs from both detectors (each output is a set of bounding boxes, each of them with an associated confidence), it is necessary to normalize both confidence outputs. Therefore, we normalize both detectors, C_{SP} ($0 \leq C_{SP} \leq 1$) and C_{WU} ($0 \leq C_{WU} \leq 1$). The normalization is performed according to the probability density function (pdf) of each Detection Confidence. In particular, the Standing People Detection Confidence

distribution has been estimated using the detector output over the INRIA dataset [Dalal and Triggs, 2005], whilst the Wheelchair Users Detection Confidence distribution is obtained detecting the wheelchair users from the training images set. Using the score histogram, the pdf is estimated trying to adjust properly to the obtained scores. The score histogram and the estimated pdf are shown in Figure 4.3.

In order to facilitate comparison between models, pdf and cdf (cumulative distribution function) are represented in Figure 4.4 for both standing people (from [García-Martín and Martínez, 2015a]) and wheelchair users models. As the considered detection algorithm is the same for both models, the density functions obtained are relatively close, but this conversion should be performed to join the detectors results rigorously. After normalizing the detections of the different models, both sets of detections are joined together to obtain the complete set that considers the different people appearances.

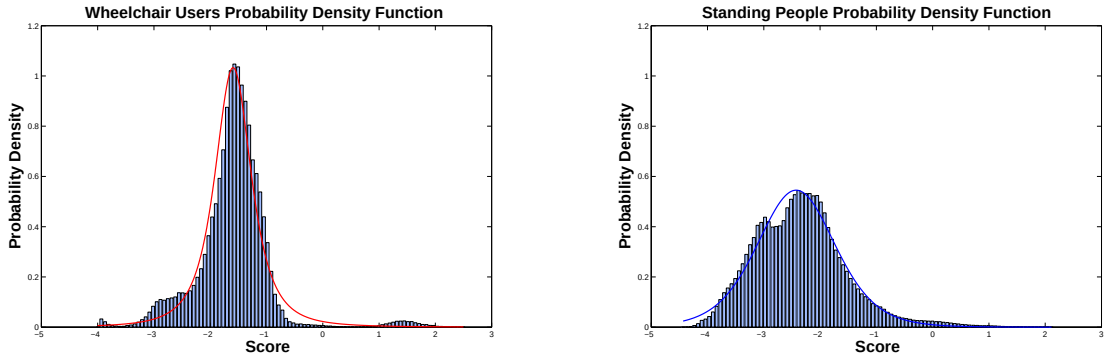


Fig. 4.3. DPM wheelchair user (left) and standing people (right, extracted from [García-Martín and Martínez, 2015a]) models score histograms with the fitted pdfs.

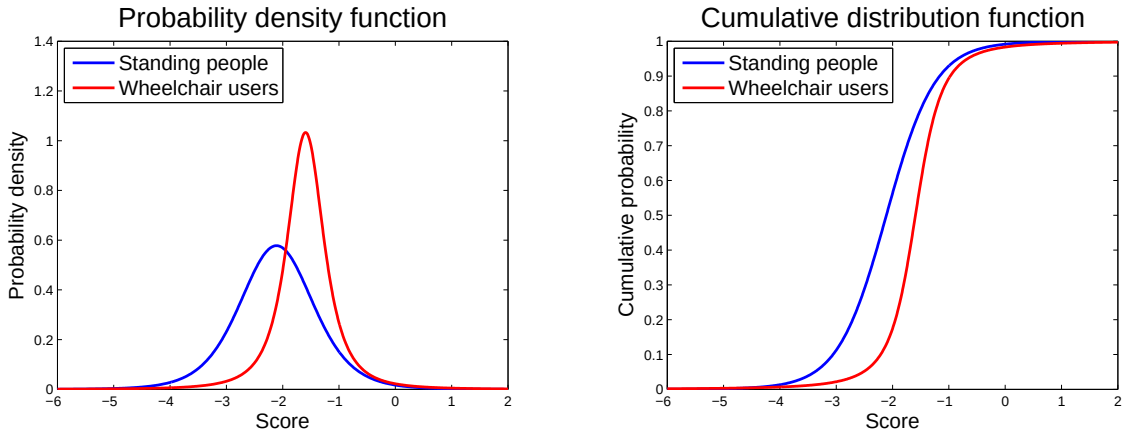


Fig. 4.4. DPM Standing people and wheelchair user detectors pdf (left) and cdf (right).

	Ground Truth All (SP+WU)			
	Smile		WUds	
	DPM	Faster-RCNN	DPM	Faster-RCNN
SP	0,777	0,734	0,577	0,688
WU	0,405	0,712	0,637	0,735
SP+WU	**0,864**	**0,777**	**0,733**	**0,811**
%Δ vs SP	11,2	5,9	27,0	17,9

Table 4.1: Detectors AUC using complete (standing people, SP, and, wheelchair users, WU) ground truth.

The Faster-RCNN output detections are by default normalized between 0 and 1 in the algorithm, so this step does not apply to its results as the normalization is internally included in the algorithm.

4.4 Experiments and results

The trained detectors (DPM and Faster-RCNN) are run on the evaluation datasets in order to analyze their performance. The SMILE wheelchair dataset [Huang et al., 2010] was used for the models generation (see Section 4.3.2) and validation (see Section 4.4.1). The Wheelchair Users dataset was used to check the generated models in a different and independent scenario. As the wheelchair users models were trained using the SMILE dataset presented in subsection 2.2.3 (using the training data), the results obtained on its test images are expected to be better than the results obtained on the images of the WUds presented in subsection 2.3.1, as it is a completely independent scenario with different wheelchairs than those used to train the model. The considered metrics for the evaluation are Precision, Recall and AUC (see Subsection 2.5.1 for more details of these metrics).

This section contains results of the detector on the Smile Lab dataset (see subsection 2.2.3) and over the Wheelchair Users dataset (see subsection 4.4.2).

4.4.1 SmileLab dataset results

Figures 4.5 and 4.6 show the resulting precision-recall detection curves obtained for the detection on the Smile Lab dataset test sequences. Table 4.1 presents the numerical AUC values of the precision-recall detection curves. All these curves are also available for downloading in the publication webpage [2].

The combination of the detection results of both models (standing person and wheelchair user models) improves the results of each model separately, for both detection algorithms (DPM

[2] http://www-vpu.eps.uam.es/publications/IncorporatingWheelchairUsersInPeopleDetection/WU.htm

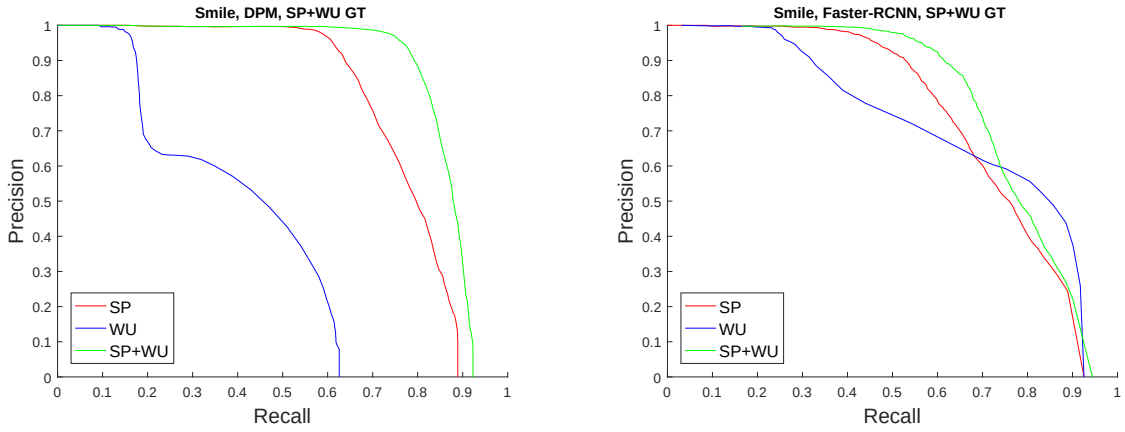


Fig. 4.5. Precision vs Recall detection curves for the Smile Lab dataset test sequences using complete (standing people, SP, and, wheelchair users, WU) ground truth.

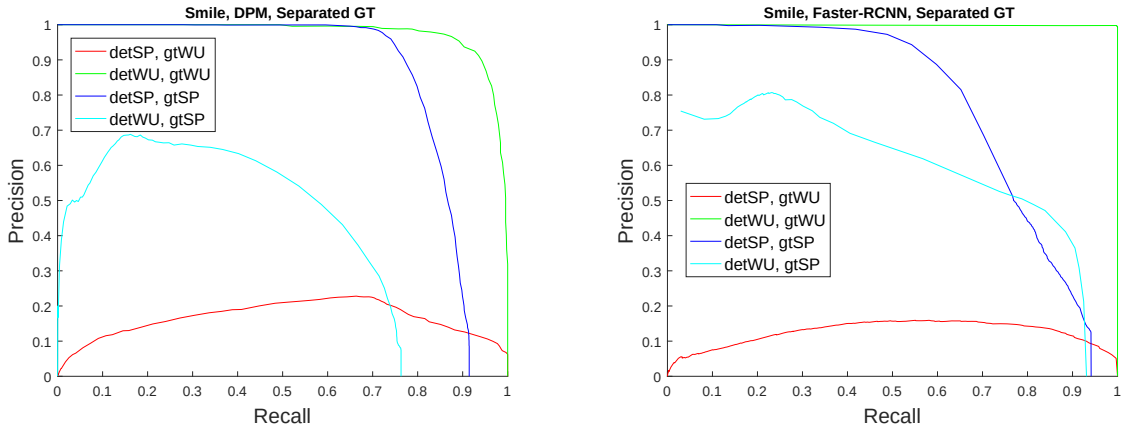


Fig. 4.6. Precision vs Recall detection curves for the Smile Lab dataset test sequences using separated detection results and separated ground truth. detSP corresponds to the Standing Person model detections, detWU corresponds to the Wheelchair Users model detections, gtSP corresponds to the Standing Person ground truth, and gtWU corresponds to the Wheelchair Users ground truth.

			Ground Truth			
			Smile		WUds	
			SP	WU	SP	WU
Detector	DPM	SP	0,852	0,163	0,883	0,283
		WU	0,415	0,977	0,265	0,833
	Faster-RCNN	SP	0,767	0,124	0,728	0,391
		WU	0,599	0,999	0,268	0,912

Table 4.2: Detectors AUC using separated detection results and separated ground truth.

and Faster-RCNN), as seen in Table 4.1, in which the area under the curve of the combination of models is better than the detection of each model separately, for both detection algorithms. The final results obtained by the DPM detector are better than those obtained by the Faster-RCNN, but it is probably due to the fact that the wheelchair user detector detects a greater number of standing people (as shown in Figure 4.6 and in Table 4.2), and its combination with the standing people detector does not manage to combine correctly in cases when there are multiple bounding boxes of the same person with a very different aspect ratio. The result of the Faster-RCNN is better when using a different dataset for evaluation (see the following section). With respect to the results using separated ground truth (Table 4.2),that is, to evaluate on the one hand the detection of standing people, and on the other hand wheelchair users, the DPM detector is able to better detect standing people, but wheelchair users are better detected by the Faster-RCNN model. The proposed detection improves the initial performance 11,2% and 5,9% on average for this dataset (see Table 4.1). Note that the training images and the test images are different but contain the same (people and wheelchair model) standing people and wheelchair users.

The obtained results can not be directly compared with the results presented in [Huang et al., 2010] for several reasons. Only the wheelchair users are detected in [Huang et al., 2010], while we detect both wheelchair users and standing people, but for this comparative we will use only the wheelchair users model detections. Also they consider a detection error when a wheelchair is detected with an orientation (among eight possible orientations) different from the one annotated in the ground truth. The work presented in this chapter does not consider the wheelchair orientation, as defined in previous sections. The authors of this dataset did not provided us the ground truth that they had used, so we had to generate a new one, as commented in subsection 2.2.3. Table 4.3 shows the results given by [Huang et al., 2010] and our wheelchair users detection results. We have selected the closest point between our precision-recall curve and the point given by the authors of the dataset. The results obtained by the DPM detector are very close to those presented in [Huang et al., 2010] but slightly worse, and the results obtained by the Faster-RCNN are significantly better, especially highlighting the null value of miss detections in all sequences. It is noteworthy that our ground truth has more frames annotated than the results given by [Huang et al., 2010] (1314 vs 1169 frames). Our ground truth has annotations of every sequences frames, regardless of it complexity, the existence of occlusions, etc. We present a different approach than [Huang et al., 2010], integrated into a complete system, but the result greatly improves the detections scores when using the Faster-RCNN detector.

4.4.2 Wheelchair Users datasets results

Figures 4.7 and 4.8 show the resulting precision-recall detection curves obtained for the detection on Wheelchair Users dataset sequences. Table 4.1 presents the numerical AUC values of the precision-recall detection curves. All the obtained curves are available in the webpage: http://

	H	M	F	R	P
[Huang et al., 2010]	1086	83	73	0.929	0.937
DPM	1218	96	99	0.927	0.925
Faster-RCNN	**1314**	**0**	**7**	**1**	**0.995**

Table 4.3: Comparative results for the wheelchair users detections between Huang et al. [2010] and our approaches. H, M, F, R and P are, respectively, hits, miss detects, false detects, recall and precision.

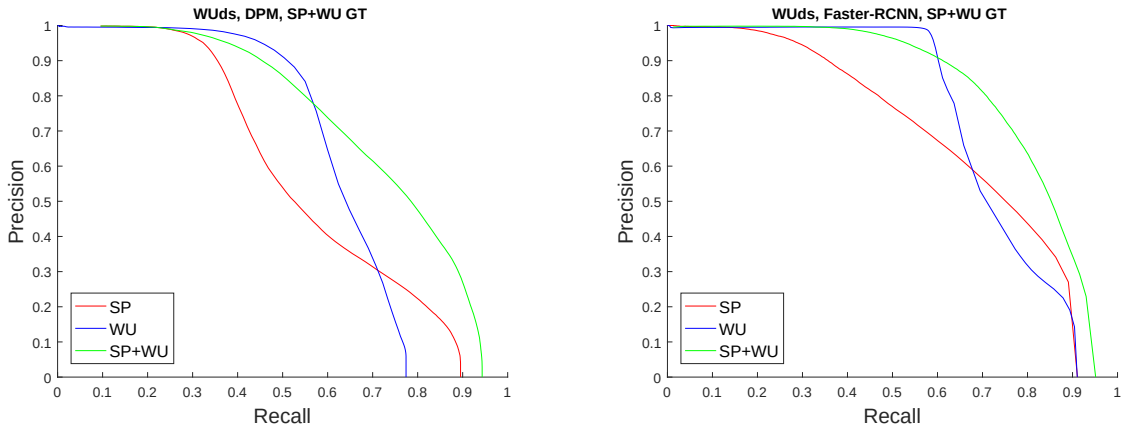


Fig. 4.7. Precision vs Recall detection curves for the Wheelchair Users dataset using complete (standing people, SP, and, wheelchair users, WU) ground truth.

www-vpu.eps.uam.es/publications/IncorporatingWheelchairUsersInPeopleDetection/WU.htm.

In the Wheelchair Users dataset sequences there is a greater number of wheelchair users, both in absolute value (greater number of wheelchair users in the sequences) and relative value (wheelchair users vs standing people ratio), so it is expected to get a greater improvement with respect to the original standing people detector. In this case, the percentage increase of the AUCs, compared to the initial detector, is 27,0% and 17,9% on average, much higher than the 11,2% and 5,9% obtained in the previous dataset (see Table 4.1). The Faster-RCNN detector performance is better in this dataset than in the one discussed in the previous section. With respect to the evaluation with partial ground truths (Table 4.2), the results obtained with the WUds are the combination of the detection results of both models (standing person and wheelchair user models) improve the results of each model separately, for both detection algorithms (DPM and Faster-RCNN), as seen in Table 4.1, in which the area under the curve of the combination of models is better than the detection of each model separately, for both detection algorithms similar to those observed with the Smile dataset.

The transfer learning to the new sequences is generic enough to improve the results, reach-

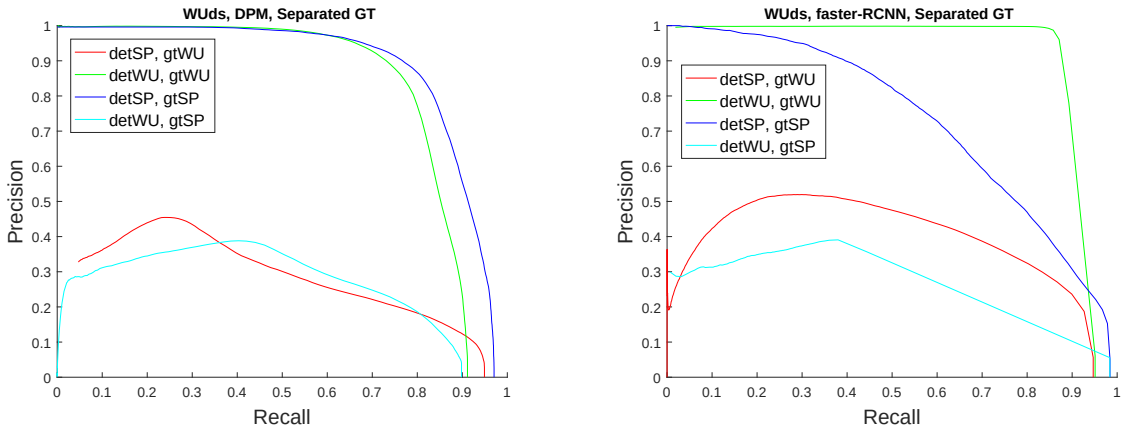


Fig. 4.8. Precision vs Recall detection curves for the Wheelchair Users dataset sequences using separated detection results and separated ground truth. detSP corresponds to the Standing Person model detections, detWU corresponds to the Wheelchair Users model detections, gtSP corresponds to the Standing Person ground truth, and gtWU corresponds to the Wheelchair Users ground truth.

ing in fact a higher percentage increase in the recorded sequences than for the Smile dataset sequences when using the Faster-RCNN algorithm. The new recorded scenario dataset presents a more realistic scenario for the detectors, where not all the wheelchair types can be considered in the model, in the same way as in the standing people detector not every person, orientation and pose are present. The recorded sequences also contains severe illumination changes and occlusions.

4.5 Nursing home map application

In addition to the detector models combination described in this chapter, a final application is shown on which the detectors output is given temporal continuity (by associating closer positions, or color histograms of detected people) obtaining a trajectory for each standing and sitting person, projecting it on the plane of a nursing home. The video used to obtain these trajectories can not be shown due to privacy issues, but the final result is shown to illustrate the final application considered. Figure 4.9 shows an example frame of the map application. In addition to trajectories, a label can be asigned to each person based on their category, for example: employee of the residence, resident, visitor.

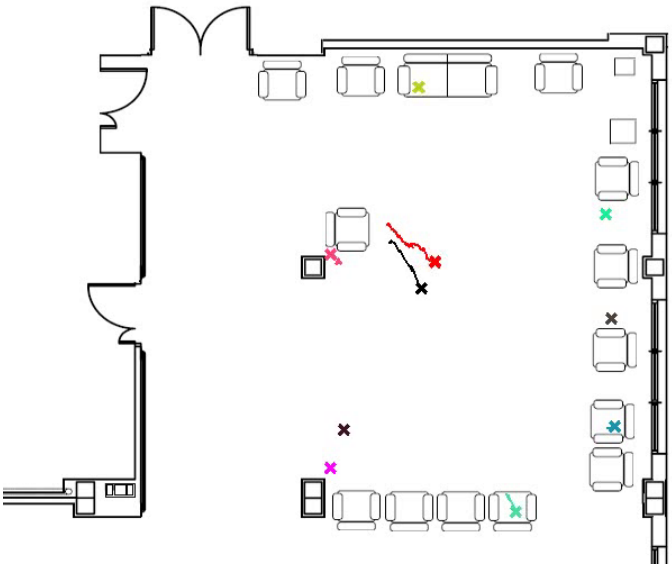


Fig. 4.9. Map application example in a nursing home.

4.6 Conclusions

In this chapter, we treat the problem of different appearances (standing people and wheelchair users) for the same semantic object class detection. Typical senior residences scenarios are an example of this situation. In particular, our main objective is to detect both standing people and wheelchair users simultaneously. For this reason, people detection has been completed with the capacity to detect people with the need of using a wheelchair. We have trained two additional wheelchair users detectors models whose detections can be combined with the detections obtained using the traditional standing people detectors models, providing generality and supplementary detection capacity. This approach can not only be applied to the case of wheelchairs but the ideas exposed here can be extrapolated to other scenarios where there are individuals with an appearance different from the standard, as Zimmer frames users or people using walking sticks.

Due to the appearance of wheelchairs, we have trained a model with two different variations (front/rear and side point of views), allowing to detect different orientations. The proposed detector does not consider the wheelchair orientation in the output, but we consider that this does not provide much information for the different applications derived from the detection. In any case, if the orientation estimation were interesting or necessary, the wheel ellipse can be located in the detection bounding box after detecting the wheelchair, following one of the existing methods.

Due to the absence of public datasets with this type of content, new sequences with greater complexity have been recorded in order to test the designed approach and to provide future

researchers with images and sequences for their experiments. We have made publicly available the generated wheelchair user models, the recorded sequences and ground truth files. We have proven the capacity of transfer learning from a training dataset to a new one completely independent.

Finally, an occupation map example has been shown in which the potential of using these techniques to monitor rooms of nursing homes is shown. Using the the spatial coordinates of the people, you can monitor the relationships between residents and the activities of each one of them, being able to detect anomalous behaviors to, for example, activate alarms.

There are multiple future work lines to improve the different proposals. About the wheelchair users detector, more complex models can be studied, for example considering more model variations. About the combination, we have chosen a simple technique, therefore it could be improved in order to optimize the combination of the different information sources. Also a new model can be trained using both the Smile Lab dataset and the recorded sequences to achieve greater generality. A tracker can be added to the sequence detection to combine the information extracted during the sequence frames giving temporal continuity to the detections. As the recorded dataset uses a multi-camera deployment, the detections obtained for each viewpoint can help to reinforce the detections from the other one (see Chapter 6). Apart from this, the typical lines of future work for object detection can be applied here. Finally, the extended people detection can be used as a starting point for multiple event detection systems, in scenarios where the presence of wheelchair users is very common, such as hospitals, healthcare centers or senior residences.

Chapter 5

Automatic vacant parking places management system using multi-camera vehicle detection

5.1 Introduction[1]

Parking lots are a widely used service where a great investment is made every year. The management of these car parks is very expensive and in many cases complex, especially in the case of those that have many places such as airports or large commercial areas. Solving this problem using computer vision promises a number of advantages over intrusive sensors like induction loops or other weight-in-motion sensors Fabian [2008]. In addition, a vision-based system may provide many value-added services, like parking space guidance and video surveillance Huang and Wang [2010]. Such systems allow the decongestion of crowded parking areas, directing vehicles to areas with lower occupancy, guiding the vehicles by a faster route.

Surveillance cameras are readily available in most car parking lots, so in many cases the solution is only to adequately process the information available from the already existing cameras, or complete the deployment by adding some cameras to have a full coverage that allows the system to operate.

The previously developed systems are mainly based on image segmentation or machine learning (SVMs, NN) over spot patches, but due to the evolution in the last years of object detection algorithms, it is possible to use the detections of these algorithms for the proper operation of automatic parking management systems. This chapter presents a multi-camera system for vehicles detection and their corresponding mapping into the parking spots of a parking lot. Approaches from the state-of-the-art, which work properly in controlled scenarios, have been validated us-

[1] This chapter is an adapted version of the publications [Martín-Nieto et al., 2017]

ing small amount of sequences and without more challenging realistic conditions (illumniation changes, different weather). On the other hand, most of them are not complete systems, but provide only parts of them, usually detectors. The proposed system has been designed for realistic scenarios considering different cases of occlussion, ilumination changes and different climatic conditions; a real scenario (the International Pittsburgh Airport parking lot) has been targeted with the condition that existing parking security cameras can be used, avoiding the deployment of new cameras or other sensors infrastructures. The system is based on existing object detectors (the results of two of them are shown) and different proposed postprocessing stages. The results clearly show that the proposed system works correctly in challenging scenarios including almost total occlusions, illumination changes and different weather conditions.

This chapter is structured as follows: after this introduction, section 5.2 presents an overview of the related work. Section 5.3 presents and describes details of the complete system and each of the blocks that compose it. Section 5.4 presents the experiments and results obtained by the system. Finally, section 5.5 describes the conclusions of the chapter and some lines of future work.

5.2 State of the art

In this section, we overview works related to the proposed automatic parking management system, which try to locate occupied/empty parking spots. We have organized all the related works in three categories taking into account the technique used for the occupied/free parking spots classification: image segmentation, machine learning (SVMs, NN, etc.) over spot patch (or patches), and vehicle detection techniques based on object detectors.

5.2.1 Image segmentation based systems

Image segmentation based systems try to differentiate, in each considered frame, between vehicles and parking spots. Background subtraction is a typical technique used in this category, where an empty image is used to subtract each frame in order to get the foreground mask (vehicles). The vehicles are extracted and then mapped to each parking spot. The most representative works included in this category are [Fabian, 2008; Huang and Wang, 2010; Wang and Hanson, 1998; Yamada and Mizuno, 2001; Lee et al., 2005; Bong et al., 2006, 2008; Lin et al., 2006; Chen et al., 2010; Blumer et al., 2012; Liu et al., 2013; Hilal Al-Kharusi, 2014; Masmoudi et al., 2014]. The algorithm from [Fabian, 2008] considers three main processing stages: firstly, shadows in the image are attenuated (or removed) and image distortion is corrected; afterwards, correspondences are established between stationary cameras and visible parking places, and, finally, the parking place status is evaluated. Status classification is based on the assumption that the surface of a vacant parking place is relatively invariant in comparison to an occupied

place. The parking slots labelling process is treated in [Huang and Wang, 2010] as a color classification process which decomposes the image observation into an object component and a lighting component. The object type is either "car" or "ground", and the lighting condition is either "shadowed" or "unshadowed" (the system needs to know the direction of sunlight). Both the expected object map and the expected shadow map are created to help in the image pixels labelling. A frame preprocessing is applied using the Surface Texture and Microstructure Extraction (STME) in [Wang and Hanson, 1998], resulting in an image where the vehicles appear as "bumps" in an elevation map. A method for individual vehicle detection using grayscale images acquired from an elevated camera is presented in Yamada and Mizuno [2001]. Vehicles are considered to be composed of several components such as hood, windows, headlights, etc., so images of parking cells are fragmented by gray level and a cell is considered occupied if it is composed by a large number of small components. A dual camera device was designed and calibrated manually in [Lee et al., 2005], where parking lots (manually specified) are detected using background subtraction. After that, two morphological operations, erosion and dilation, are performed to connect the blobs and to eliminate the noise. The system introduced in [Bong et al., 2006], and enhanced in [Bong et al., 2008], needs to store an unedited zero occupancy image, and manually store the identified coordinates of every parking spot. The object (vehicle) detection is based on a combination of background extraction and edge detection (using the Sobel operator). Background subtraction is also used in [Lin et al., 2006] but with two additional considerations: a preprocessing color filter is applied for maintaining color stability, and a shadow removal is used to remove shadow foreground pixels. A spot is considered occupied if the percentage of the foreground pixels in the spot patch are over an empirical threshold. A stitching algorithm is used in [Chen et al., 2010] to integrate visual cues from multiple cameras for constructing a panoramic scene. Color, position and motion are used for tracking vehicles across different cameras. Two features are used to capture the vacant properties of each parking space: edge (Canny filter) and color (background subtraction). Three different methods of image analysis are combined in [Blumer et al., 2012]. Edge counting and histogram classification are utilized as static analysis methods (information available in a single frame) and a crafted algorithm for blob tracking as dynamic (across-frames) method using background/foreground estimation. The occlusion problem, which is important in other approaches, is supposed to be avoided in the paper through camera placement at high floors, which is not always a possible solution. In [Liu et al., 2013], after an initial edge detection stage, edge density, closed contour density and foreground/background pixel ratio are combined to decide whether a car is present or not in each parking spot. The parking space boundaries are fixed, and the region of each parking space can be defined using 4 dots or just a parallelogram as a given parking spot. After that, each parking space is numbered. The parking management system described in [Hilal Al-Kharusi, 2014] tries to find the car park coordinates from an empty car spot, acquiring an

image with cars, converting the image to black and white for simple analysis, removing noise and determining whether car spots are vacant or filled. Each spot is segmented to decide whether it belongs to the background (empty) or to the foreground (occupied). Two types of car parking lots photos are used: one is taken from Google earth and the other one is a real car park photo. After a homography transformation, the system presented in [Masmoudi et al., 2014] performs a background subtraction, and a feature classification (SURF [Bay et al., 2008] and HOG [Dalal and Triggs, 2005]) to decide the status of each parking spot.

5.2.2 Spots Patch classification based systems

Spots patch classification based systems use classification machine learning techniques (SVM, NN, etc.) which are trained with previously labelled patches of occupied and free parking spots. The most representative works included in this category are [Sastre et al., 2007; True, 2007; Wu et al., 2007; Huang et al., 2008; Al-Absi et al., 2010; Huang et al., 2012, 2013a, 2015; Tschentscher et al., 2015; Huang and Vu, 2015]. The parking management system presented in [Sastre et al., 2007] creates, using homography computation, a pseudo-top-view of a parking area to determine if there are free parking lots or not. The texture feature extraction of each parking lot is obtained using Gabor filter banks. A SVM is trained with texture feature vectors of every parking spot, which have been taken in different illumination conditions and with diverse type of shadows. The algorithm proposed in [True, 2007] uses a combination of car feature point detection and color histogram classification to detect vacant parking spots. The author points out that one major weakness of this algorithm is that it can not accurately detect the state of parking spots which are slightly or mostly occluded by objects such as other vehicles. The method for parking space detection proposed in [Wu et al., 2007] trains and recognizes empty parking spaces by applying machine learning methods (SVM). Three consecutive parking spots are proposed as a detection patch, which contains the space under consideration and the two neighboring spaces. The system uses PCA to pick 50 critical features. The problem is addressed in [Huang et al., 2008] through a Bayesian hierarchical detection framework. The top layer is an observation layer, where each node indicates a local feature. The local feature can be either texture-based or pixel-based . Haar-like features are used in [Al-Absi et al., 2010] for the detection of features detected in input videos to determine the presence of a car within a parking spot. A surface-based hierarchical framework is proposed in [Huang et al., 2012] to integrate the 3-D scene information with the patch-based image observation for the inference of vacant space. The HOG feature dimension is reduced using a Linear Discriminant Analysis (LDA), and 4 likelihood models are trained for each surface type. A classification of several algorithms for vacant parking space detection is presented in [Huang et al., 2013a], depending on the challenges that they consider for vacant parking space detection: perspective distortion, inter-object occlusion, shadow effect, lighting variations and insufficient illumination at night.

They use HOG features for car detections, and cars are decomposed into four types of planar surfaces. Since the perspective projection process is highly dependent to the camera setting, the patch classification models need to be re-trained for different camera settings. It takes two days to install the system. The first day is used for hardware setup, camera calibration and training data collection. The second day is used to label the training data and to learn models for patch classification. Most of the failure cases are caused by the headlight of moving cars. [Huang et al., 2015] extends the system presented in [Huang et al., 2013a] adding a multiclass boosting method to automatically select the weak classifiers weights through a back-propagation learning process. This system is divided into 3 layers: 3D-cuboid model and feature extraction layer, patch classifier layer, and weighted combination layer. Like in other systems already mentioned (e.g., [Huang et al., 2012]), a LDA process is used to reduce the feature dimension of the extracted HOG features. In [Tschentscher et al., 2015], several features with different color histograms or DoG histograms are analyzed using three supervised learning algorithms (k-NN, LDA, SVM). Finally, a multi-layer discriminative framework for vacant parking space detection is presented in [Huang and Vu, 2015]. This extended framework adds a status inference layer over [Huang et al., 2013a].

5.2.3 Object (vehicle) detectors based systems

Object (vehicle) detectors based systems use a detection algorithm to detect vehicles and to map them into the different parking spots. This type of system has begun to be viable in recent years thanks to the evolution of object detectors, specifically [Girshick et al., 2013; Girshick, 2015; Ren et al., 2015]. The only workto our knowledge, included in this category is a car detection method [Xie et al., 2015] based on the Convolutional Neural Networks (CNN) technology. After training the CNN, to identify where there are cars, they search the whole image of a parking lot using a sliding window approach. In this work, the detection is performed but the results are not mapped in the different parking spots and, therefore, it is not a complete system.

In our work, we also propose to follow the detection approach but designing and developing the different stages to get a complete automatic parking management system: vehicle detection, homographic transformation, perspective correction (for allowing to reuse existing camera instalations), automatic spot mapping and multi-camera fusion (assuming the usual avaliability of multi-camera setups). Additionally we have created a complete realistic dataset including a multi-camera environment with both illumination and climate variability and we perform a rigorous and methodological evaluation of the proposed system.

5.2.4 Qualitative comparison between existing approaches and the proposed system

Due to the absence of public datasets of stationary vehicles, it is not possible to make a quantitative comparison of the proposed detection based system with respect to the others, however, the novelty of the proposed detection based system allows to conceptually compare the advantages of the system compared to the existing ones.

An advantage of the proposed system over existing systems is the "automatic vehicle mapping" on the different parking spaces. Many approaches (e.g., [Bong et al., 2008; Lin et al., 2006; Hilal Al-Kharusi, 2014; Sastre et al., 2007; True, 2007; Al-Absi et al., 2010]) require manually annotating, one by one, the position of each spot, while our system needs only the corners of the parking area and the number of spots. This advantage is especially notable compared with the spots patch classification based systems and especially in the case of large car parking in which the number of places to label is high. The main advantage of the proposed system over the image segmentation based systems is the robustness against variable background, generally caused by climatic or lighting variability. This system is the first of its class to detect and subsequently map in the different parking spaces, as Xie et al. [2015] just detects the vehicles and does not perform the subsequent steps.

Another advantage of detection based systems is the capacity to withstand "object occlusions". Although some of the existing systems (e.g., [Wu et al., 2007; Huang et al., 2013a, 2015; Huang and Vu, 2015]) already try to support occlusions, the object detectors have a better capacity to support them because they use the information they have without needing to add dependency occupation rules between adjacent spots.

Finally, adding "multi-camera support" to the system allows the existence of complete occlusions in the scenario, and the use of redundant information from the different cameras allows to improve the system performance.

5.3 Proposed system

5.3.1 Overview

The proposed multi-camera system is based on a parallel processing of each camera followed by the combination (or fusion) of their individual results. The block diagram of the system is presented in Figure 5.1. Each camera captures frames, which are processed frame-by-frame. Firstly, an "object detector" (using a previously trained vehicle model) locates the vehicles in the frame; using an "homographic conversion" and a "perspective correction" to consider the volume of the detected objects, the obtained detections are "automatically mapped" into the positions of the occupied/empty mono-camera spot matrix. Finally, if there is a mutlicamera

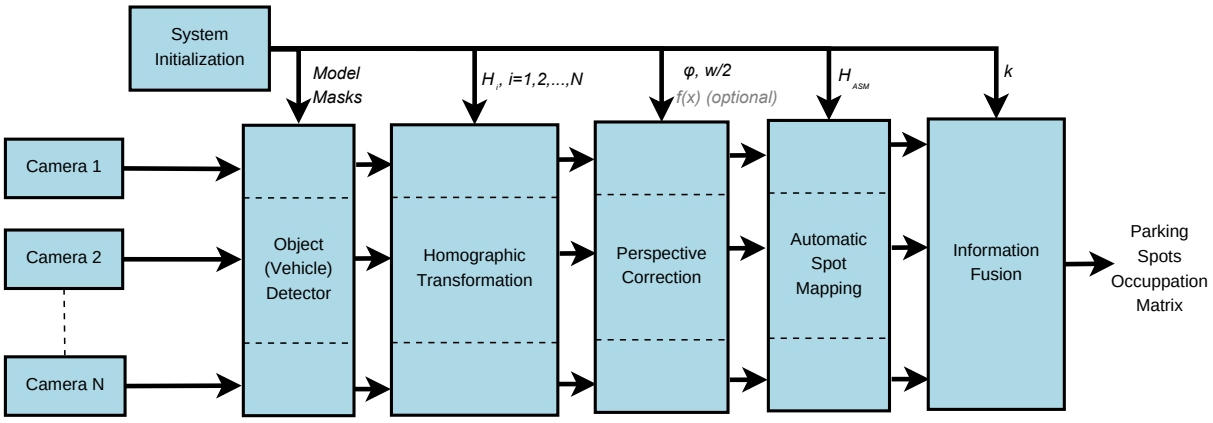


Fig. 5.1. System Block Diagram: cameras provide the frames to be processed (left); the initialization block provides the necessary information for each block (top); and the result of the system is the parking spots occupation matrix (right).

setup, the information from each camera is "fused" to obtain the final multi-camera spot matrix which indicates the occupation of the parking lot.

In order to present a system configuration example, we proceed to describe the source of each of the modules of our implementation: the considered detectors are existing techniques from the state of the art but their models have been trained specifically for the purposes of the system; the homographic transformations are mathematical techniques described in [Hartley and Zisserman, 2003] but we have generated our own homography matrix for each of the cameras; the perspective correction is a technique designed by us for this system and is based on trigonometry; the automatic spots mapping is a technique designed by us based on homographies; the fusion considers tuned functions of standardized sigmoids.

5.3.2 Object (vehicle) detector

The object vehicle detector is initialized with the vehicle model, and, in order to eliminate possible detections of other areas of the parking lot that will not be monitored with these cameras, it also receives a region of interest (ROI) mask for each camera. An example of these masks is shown in Figure 5.2.

This block receives the frames of each camera and, using an object detection algorithm, generates as output a bounding box (rectangle) for each of the detected objects (vehicles).

We have trained new vehicle models because existing car models do not function properly when using an image with a high viewpoint, scales variability, occlusions, different vehicle types, etc., as contemplated in the experiments sequences.

The main detection algorithm seleceted for the evaluation of the proposed system is the Faster R-CNN (Regions with Convolutional Neural Network Features) [Ren et al., 2015] detector, which

a b c

Fig. 5.2. Example of (a) ROI mask, (b) input frame and (c) masked frame.

is a more efficient variation, mainly in terms of computational cost but also in performance, of the previous R-CNN [Girshick et al., 2013] and Fast R-CNN [Girshick, 2015] detectors. The three variations have in common the combination of bottom-up region proposals with rich features computed by a convolutional neural network. The main difference of the Faster-RCNN is the use of a Region Proposal Network (RPN) that enables nearly cost-free region proposals. For training purposes and according to the author's results [Ren et al., 2015], we have chosen the pre-trained network VGG-16 model [Simonyan and Zisserman, 2014] that has 13 convolutional layers and 3 fully-connected layers. We have used the network to train a new model using the PASCAL VOC 2007 and 2012 datasets and we have added a new object class, our parking vehicle model, using our dataset (see section 2.3.2). The vehicle detector used in [Xie et al., 2015] is also based on a generic CNN from the state of the art, trained by the authors. As the code and model are not publicly available, it can not be used in the evaluation of the system, but it could be integrated and evaluated in a direct way

The second detection algorithm evaluation is the Deformable Parts Model (DPM) detector [Felzenszwalb et al., 2010b]. The DPM detector is based on exhaustive search and a part-based model. It is a part-based adaptation of the original Histogram of Oriented Gradients detector (HOG) [Dalal and Triggs, 2005]. It proposes an object detection system based on mixtures of multiscale deformable part models where each deformable object part is modeled as the original HOG detector [Dalal and Triggs, 2005]. The algorithm model also contains the flip of the model. We used this detector in order to see the behavior of the system when using a non deep-learning based detector. As deep-learning based ones are "better" detectors, this evaluation allows to demonstrate the robustness of the system to detection noise.

Additionally, experiments were also made with the ACF (Aggregate Channel Features) Dollar et al. [2014] algorithm, but, due to the properties of this technique, the bounding boxes obtained during the detection process covered only the roof of the vehicles, instead of completely covering them. This causes that this algorithm does not fulfill the requirements for the detectors of the proposed system which considers that the bounding boxes completely cover the vehicle.

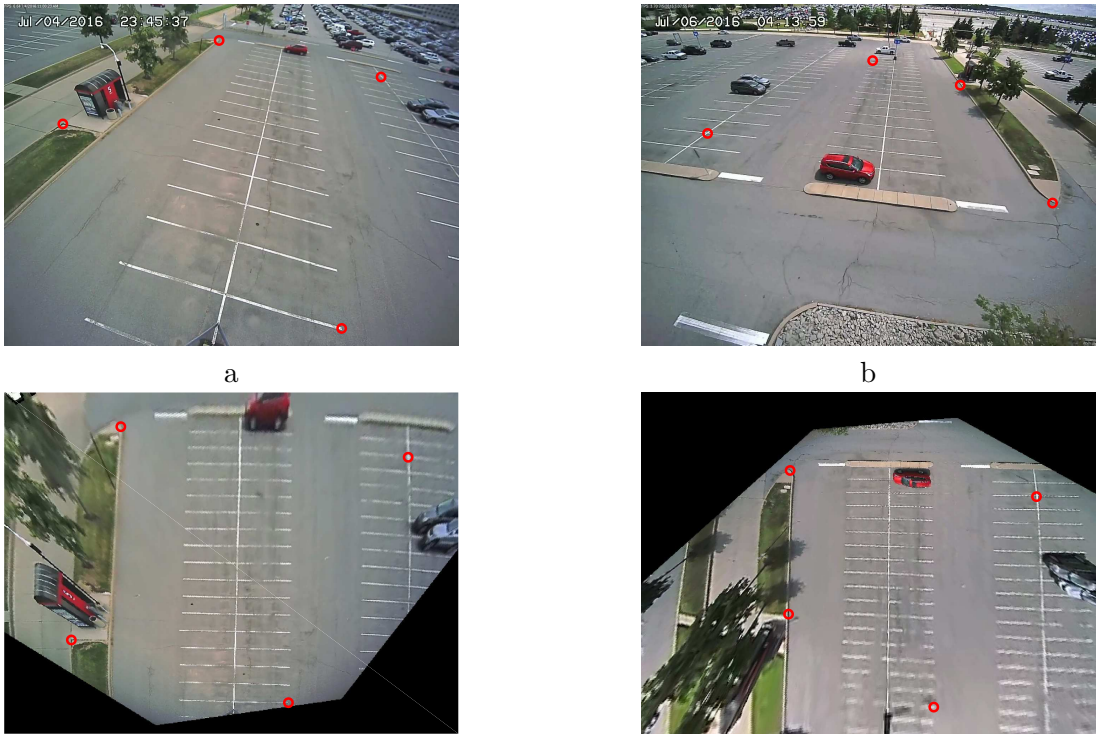


Fig. 5.3. Homography viewpoint transformations: (a) and (b) show the starting side viewpoints, (c) and (d) show the resulting top view common planes.

5.3.3 Homographic transformation

The object detector of the previous block obtains a bounding box for each detected vehicle from the viewpoint plane of each camera. This block, using the properties of the homographies, allows to change the position of the objects detected from the plane of each camera to a common plane. The homography matrix (which is needed to initialize the block), H_i, for each camera i, is obtained using 4 points from each camera viewpoint and each point correspondence in an image extracted from a top view. This top view can be easily obtained from *Google Earth*. It is not necessary to choose the same points in each camera viewpoint for all the cameras, but each selected point must be associated with one from the image of the common ground plane. Figure 5.3 shows two examples of the resulting viewpoint change using homographies. Note that these images are generated only to illustrate the procedure, but this computationally expensive step is not required during the system operation by the mapping algorithm. Therefore, homography is just applied to the base midpoint of each bounding box resulting in an optimized computation. The output of this block is one point for each detected vehicle.

5.3.4 Perspective correction

Due to the volume of the detected objects, it is necessary to correct the positions of the projected points where the detections, received from the previous stages, are mapped. It is possible to make a position correction using the angle between the parallel lines of the parking spaces and the camera viewpoint. This allows the correct matching between the vehicle detections and the parking spots. Figure 5.4 presents the correction diagram and an example. In the diagram, A corresponds to the base midpoint detection projection, B corresponds to the final position after correction, φ is the angle between the parking lines and the camera view (needed for the block initialization), and $\frac{w}{2}$ is the half of a vehicle length (average). Despite referring to the length of the vehicle, the letter w is used to associate it with the width of the bounding boxes. In the example, the blue line in Figure 5.4(b) represents the base line projection of the detection bounding box. Note that the midpoint of the base is a different point than the center point of the bounding box; the midpoint of the base, belonging to the ground plane, allows the fulfillment of the properties of the applied homography. In addition, the choice of this point allows the system to work independently of the height at which the cameras are located, as its projection is always placed between the vehicle and the camera (see Figure 5.4(b)).

On the other hand, it is possible that the distortion of the lens affects the accuracy of the homography, which causes errors and imprecision in the mapping of the spots, as seen in the Figure 5.5(a). This problem is usually solved using the intrinsic camera parameters. If these parameters are not available, there is an alternative solution that consists of correcting the mapping of the grid of spots using a simple linear adjustment function (Figure 5.5(c) shows an example of adjustment function). Figure 5.5(b) shows the result of applying this correction. We define a uniform grid and then we add a correction factor to the projected point in order to eliminate the effect of the lens. You can correct the grid to fit the image, or correct the image to fit the grid. For the system, it is more efficient to correct the image to fit the grid, since it allows to automate the subsequent steps without needing any other correction. The grid could also be modified, but the image correction simplifies the next step of automatic mapping, as the uniform grid allows the automatic spot mapping via homographic techniques.

5.3.5 Automatic spot mapping

This block is based on using the properties of the homographies. However, in this case the selected destination points are designed specially to get the automatic discrete spot numbers directly without the need of supervision and without the need to map each position one by one like most of the state of the art systems (e.g. [Bong et al., 2008; Lin et al., 2006; Hilal Al-Kharusi, 2014; Sastre et al., 2007; True, 2007; Al-Absi et al., 2010]). The source points are the four corners of the parking grid, and the destination points are the corners of the synthetic destination space shown in Figure 5.6, specifically: $(1,1)$, $(1,M+1)$, $(N+1,M+1)$ and $(N+1,1)$. M is the

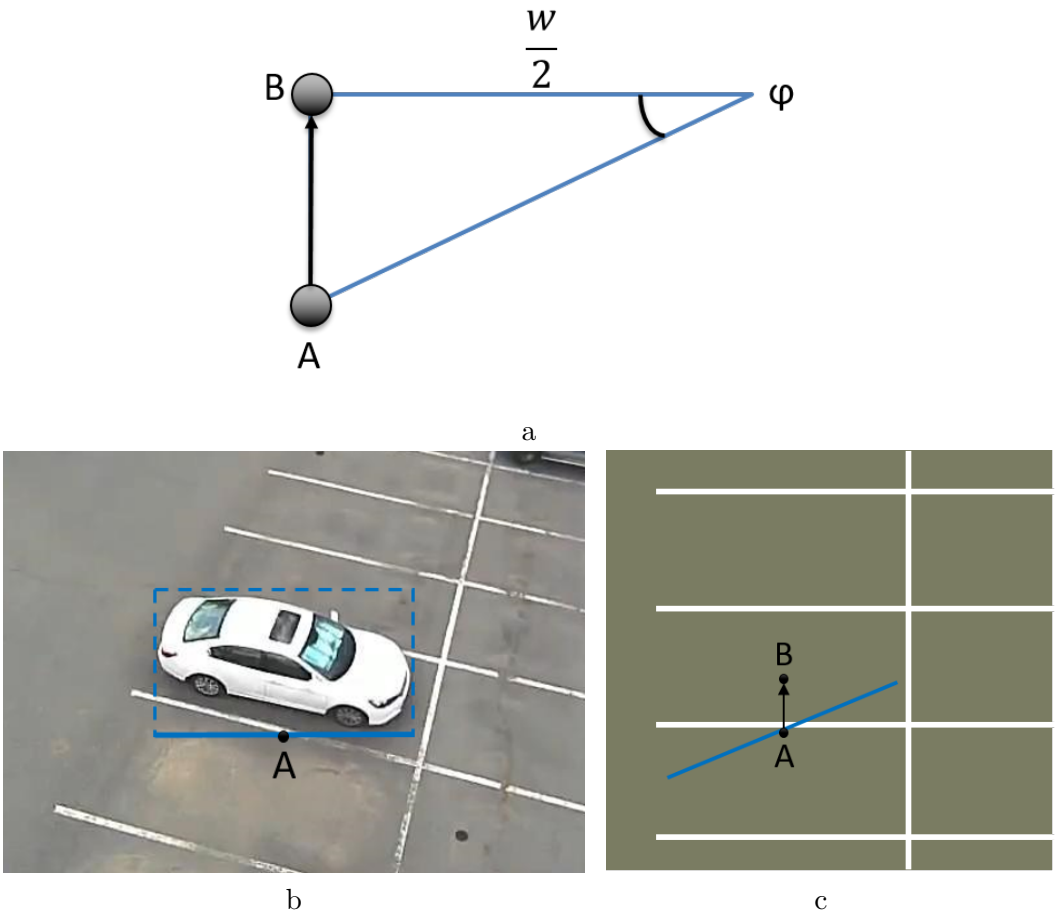
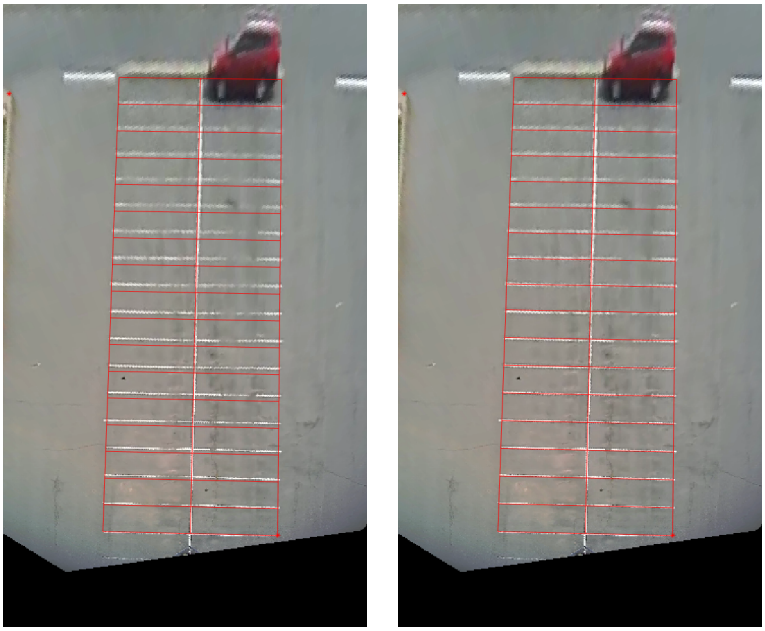
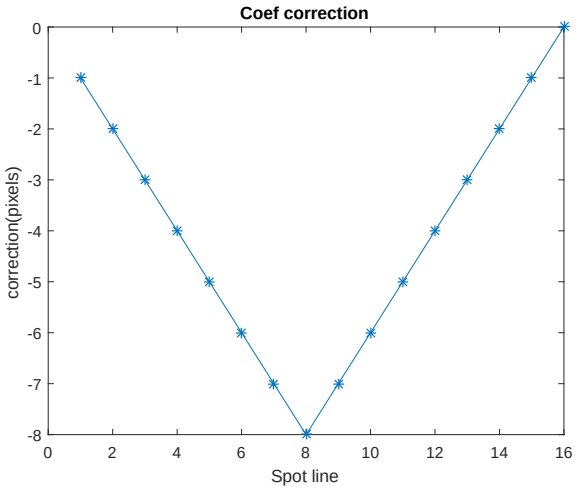


Fig. 5.4. Perspective correction diagram and example: (a) schematic diagram; (b) camera viewpoint detection; and (c) position correction example.



a b



c

Fig. 5.5. Camera lens correction: (a) initial grid, (b) corrected grid and (c) correction function.

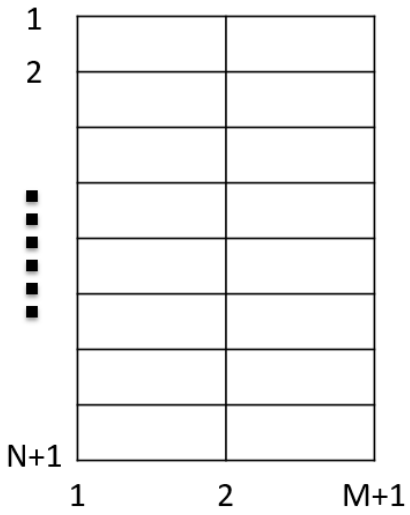


Fig. 5.6. Destination points for the automatic spot mapping.

number of parking columns (1 or 2), and N is the number of parking rows. The procedure to obtain the discrete (d) position of the mapped detection in the occupation spot matrix (x_d, y_d) consists of two steps and is presented below:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = H_{ASM} \begin{bmatrix} x_{cp} \\ y_{cp} \\ 1 \end{bmatrix} \tag{5.1}$$

where, H_{ASM} is the homography matrix for automatic spot mapping, obtained with the previously defined source and destination points, x_{cp} and y_{cp} are the x and y coordinates of the projected (and corrected in the previous stage) detections mapped in the common plane. This block needs H_{ASM} for its initialization.

Finally, the operation that allows obtaining the discrete value of the occupied spot is:

$$(x_d, y_d) = \left\lfloor \frac{x'}{z'}, \frac{y'}{z'} \right\rfloor \tag{5.2}$$

The outputs of this block are the occupancy spot matrices generated by each camera.

5.3.6 Information fusion

Logically, due to the resolution of cameras, optics, etc., the greater the distance between the camera and the mapped detections, the lower the accuracy. In order to deal with this factor, it has been decided to study a method to fuse the information of all cameras using a normalized

sigmoid function that allows to evaluate/study different combination approaches in a simple way (using a unique parameter). For this purpose, a normalized sigmoid function, $P(x)$, has been used:

$$P(x) = \frac{kx}{k - x + 1} \tag{5.3}$$

where, k is the parameter which allows to tune the sigmoid. The formula presented works for $0 < x \leq 1$, the normalized distance between the camera and the center point of the parking area. It is necessary to repeat the function for negative values, to get the range from -1 to +1. This is achieved by giving the function the absolute value of x, and then changing the sign of the result back to the same sign as x. Additionally, the result is rescaled so that at the beginning and at the end (x = 1 and x = -1) the function takes values of 1 or 0. The final normalized function, $P_{norm}(x)$, is defined as:

$$P_{norm} = \begin{cases} \frac{0.5kx}{k-x+1} + 0.5 & 0 < x \leq 1 \\ 0.5 & x = 0 \\ \frac{-0.5k|x|}{k-|x|+1} + 0.5 & -1 \leq x < 0 \end{cases} \tag{5.4}$$

Some resulting sigmoid functions are shown in Figure 5.7 with different examples for the k parameter. In this way, the camera whose detections are weighted is placed at point x = 1, and the center point of the monitored parking area is placed at point x = 0. In the case of systems with two cameras, the other camera is located at the point x = -1, but this weighting of the detection confidence does not require the system to use only two cameras since it supports any number of them. In a scenario with more than two cameras, it is necessary to define the center of all of them, and each camera will have an associated function $P_{norm}(x)$, adapted by its corresponding distance to the center.

As shown in Figure 5.7, it is possible to obtain the extreme cases in which a plane function ($k = 0$) is obtained with a constant value equal to 0.5 (all cameras have the same confidence for all the points of the parking), a step function ($k = -1$), and a straight line of slope 1 between $x = -1$ and $x = 1$ ($k = \infty$, only the nearest camera detections are considered). It is also possible to obtain symmetric curves with respect to the straight line of slope 1 (values 0.1 with -1.1, and 0.5 with -1.5).

Thanks to this confidence weighting, the most distant detections will lose score against those close to the camera. After this, the detections of all the cameras are added and are used to obtain the final parking spaces occupancy matrix. For the sigmoid with k = 0, the result is equivalent to adding all detections of all cameras with their original score, since all of them are weighted by a value of 0.5. In the case of k = -1, the only detections that are maintained in each camera are those of which the camera that detects is the closest of all of the cameras. This

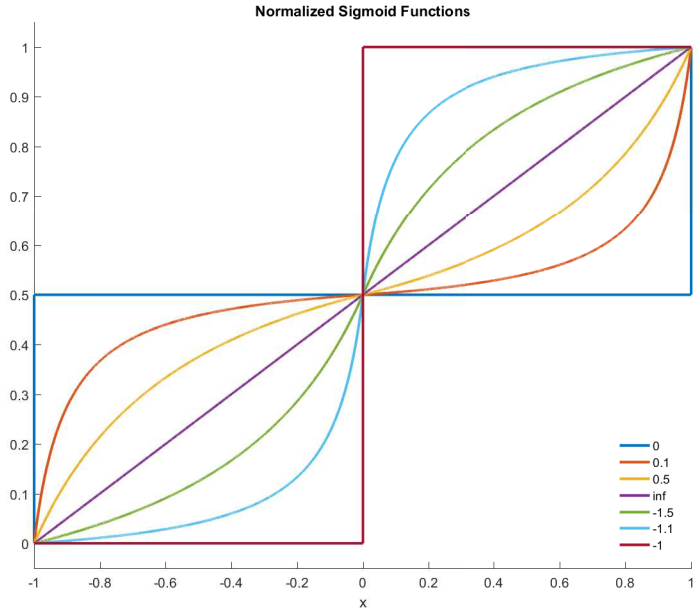


Fig. 5.7. Normalized sigmoid functions using different k parameter values.

case is a combination of information between the different cameras, each covering the area which contains the nearest spots.

For the selection of parameter k, the chosen detection algorithm and the scenario (mainly location of each camera) must be taken into account. Negative values should be considered for parameter k (e.g., -1, -1.1, -1.5) if the performance of the detection algorithm falls significantly with distance, or if the considered camera has low resolution, which complicates its detection. Otherwise, positive values of the parameter k (e.g., 1, 1.1, 1.5) will produce a better performance of the system as it considers the farther detections of each camera with greater weight.

After the automatic spot mapping (see Section 5.3.5), the occupation matrix $O_{k,i}$ for camera i and a learned k parameter is defined as:

$$O_{k,i}(x_d, y_d) = \begin{cases} 1 & if\ spot\ (x_d, y_d)\ is\ occupied \\ 0 & otherwise \end{cases} \tag{5.5}$$

One occupation matrix is obtained for each camera. All of them are fused to obtain the total occupation matrix of the system, $O_{k,T}$:

$$O_{k,T} = \bigcup_{i=1}^{n_{cameras}} O_{k,i} \tag{5.6}$$

Figure 5.8 presents a simplified example of the complete process, in order to clarify the

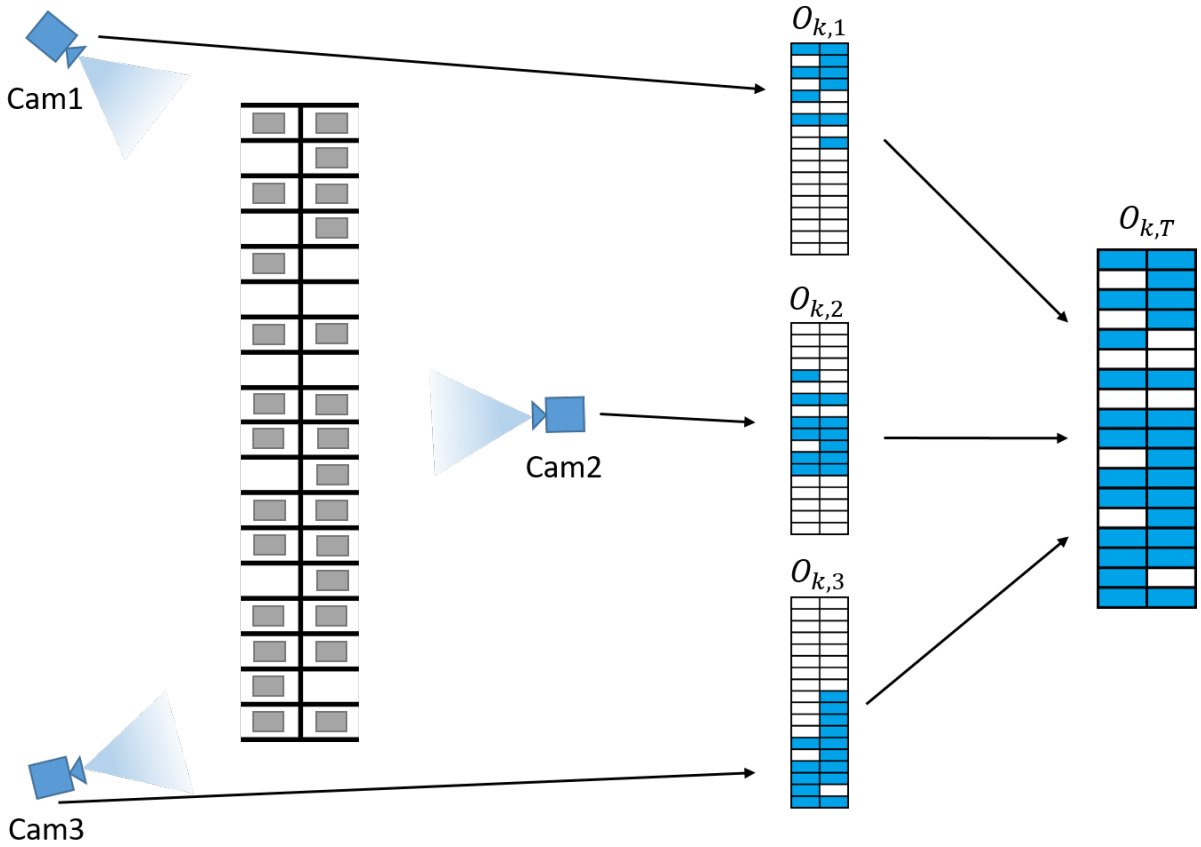


Fig. 5.8. Information fusion example: parking and monitoring cameras (left), mono-camera extracted occupation information (center) and result of the information combination (right).

information fusion stage. In the left side, there is an example of parking with occupation, and three cameras monitoring the area of interest. In the center, the occupation information extracted by each camera is processed and the mono-camera occupation matrix is generated. The right part of the example presents the result of the information combination, obtained by combining the information from all cameras.

5.4 Experiments and results

5.4.1 Detection level evaluation

As commented in section 5.3.2, the first stage is performed by the vehicle detector.

We used the Parking Lot dataset (PLds, see subsection 2.3.2) to test the designed system. We evaluate the detection results (see subsection 2.5.1 for details) with and without the use of the ROI mask (see Figure 5.2). The detection results evaluation is done using the two detection

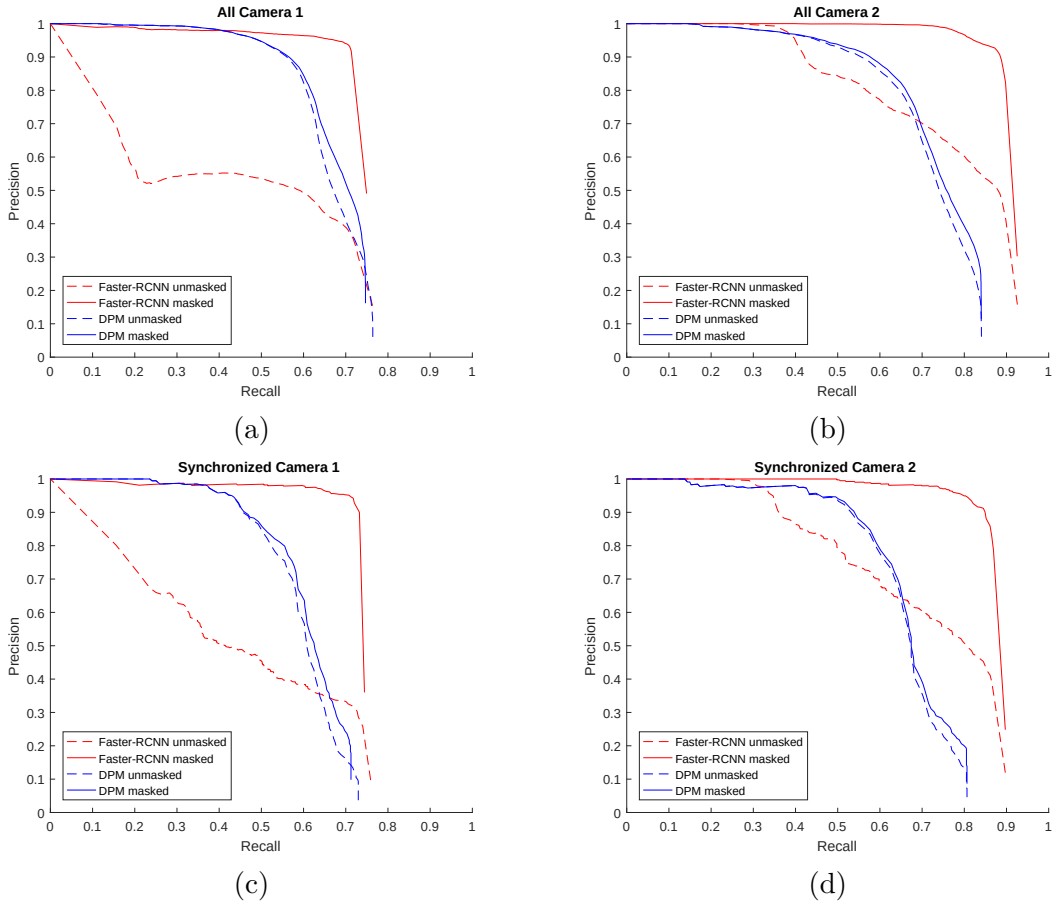


Fig. 5.9. Detection level evaluation for the two object detection trained models (Faster-RCNN Ren et al. [2015] and DPM [Felzenszwalb et al., 2010b]).

algorithms presented in subsection 5.3.2. All the generated models are executed over the four test image sets. Both detectors are evaluated with and without masking, e.g. Faster-RCNN default vs Faster-RCNN masked, in order to show its usefulness. The PR curves of this initial evaluation are shown in Figure 5.9 and the AUC values are shown in Table 5.1.

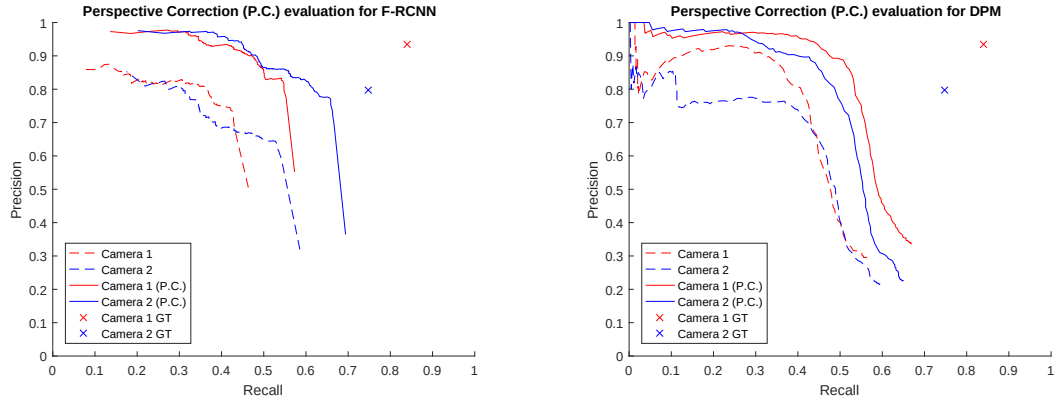
All results of the masked detector are above those of the detections without masking. In particular, the Faster-RCNN detector is able to detect vehicles from other rows not controlled by the system and, therefore, not included in the manually annotated ground truth. This stands out for the camera 1, in which precision quickly falls as the recall increases.

If we compare the performance between the two cameras, in the case of the camera 1 the Recall performance usually decreases faster since it is positioned at a greater distance from the parking area controlled by the system. This will be taken into account in later stages by combining the information from the different cameras (see section 5.4.3).

The All and Synchronized curves behave similarly, so the selection of synchronized frames is

Algorithm		All Cam1	All Cam2	Syn. Cam1	Syn. Cam2
Faster-RCNN	Unmasked	0.436	0.766	0.438	0.708
	Masked	**0.723**	**0.905**	**0.726**	**0.871**
DPM	Unmasked	0.664	0.717	0.598	0.661
	Masked	0.674	0.730	0.610	0.670

Table 5.1: AUC detection scores for detection level evaluation. The best results obtained for each image set are shown in bold.



(a) (b)

Fig. 5.10. Mono-camera spots evaluation: perspective correction evaluation for the two object detection trained models (Faster-RCNN left, DPM right) and for the ideal detection (detection ground truth).

sufficiently representative for later experiments.

5.4.2 Perspective correction evaluation (mono-camera)

The second performed evaluation considers the perspective correction. This evaluation is carried out at parking spots level with the aim of demonstrating the need to correct the projection of the detections due to the point of view. Figure 5.10 shows the improvement of performing the perspective correction for both detectors, for each of the cameras, and including the corresponding precision-recall values obtained from the use of the Ground-Truth detection. Table 5.2 presents the values of area under the curve (AUC) for all the curves shown in Figure 5.10. The results show that perspective correction is necessary and results in a significant improvement. From the Ground-Truth detection, the scores for each camera are also obtained (red and blue cross), which allow to have a measure of the best score that the mono-camera detections can reach. Despite the improvement, the results can be further enhanced by the multi-camera information combination, which is presented in the following subsection.

Algorithm		Syn. Cam1	Syn. Cam2
Faster-RCNN	Uncorrected perspective	0.380	0.452
	Corrected	0.526	**0.622**
DPM	Uncorrected perspective	0.440	0.402
	Corrected	**0.578**	0.537

Table 5.2: Mono-camera AUC scores for perspective correction at parking spots level evaluation, for the two object detection trained models and for the two cameras image subsets. The best results obtained for each image set are shown in bold.

5.4.3 Multi-camera information fusion level evaluation

Finally, using the complete system, the matrix of occupied and empty spots is obtained and it is evaluated by comparing different results (the ones for each threshold of each detector model and the parameter k) with the ground truth of the matrix of occupancy of parking spaces (see Subsection 2.6.2 for details of this specific evaluation). These results are shown in Figure 5.11. The Area Under the Curve of each detector is also computed and shown in Table 5.3. This table also contains the result of the spots evaluation using the detections ground truth (manual annotations of bounding boxes for each camera view). Despite the use of ideal vehicle detections in this case, the result of the parking spots evaluation is not perfect (the value [1,1] is not reached for precision-recall) but the impact on the score is minimal (<0.03 precision lose). This error is due to the ideal annotations of vehicles contain subjective errors of the manual annotation (e.g., object bounding box estimation due to occlusions). We consider that it is not worth trying to correct it since it allows analyzing the impact on the result, which is despicable compared to the AUC values obtained by the considered detectors. The best result is obtained with the Faster-RCNN detector, followed very closely by the DPM detector, but in all cases the results obtained are good enough for the proper functioning of the system (AUC around 0.9). Although the DPM detector presents worse results in detection (see section 5.4.1), the complete system obtains, at spots level evaluation, very close results to those obtained by the Faster-RCNN detector. It is worth to point out the difference between the results of Figure 5.3, which shows the spots evaluation at multi-camera level, with the results of Figure 5.10, which shows the spots evaluation at mono-camera level.

With respect to the different functions of normalized sigmoids used for the information combination/fusion, the results between them are very similar, but thanks to them it is possible to slightly improve the overall result of the system for a very small cost (simply weighting the detection scores of each bounding box depending on the distance to the detecting camera).

In order to configure a deployed system, the configuration of the parameters could be done in a convenient and simple way, as for the evaluation of spots it is necessary only to annotate a binary matrix of occupation (0 or 1 depending on whether the spot is empty or occupied). After

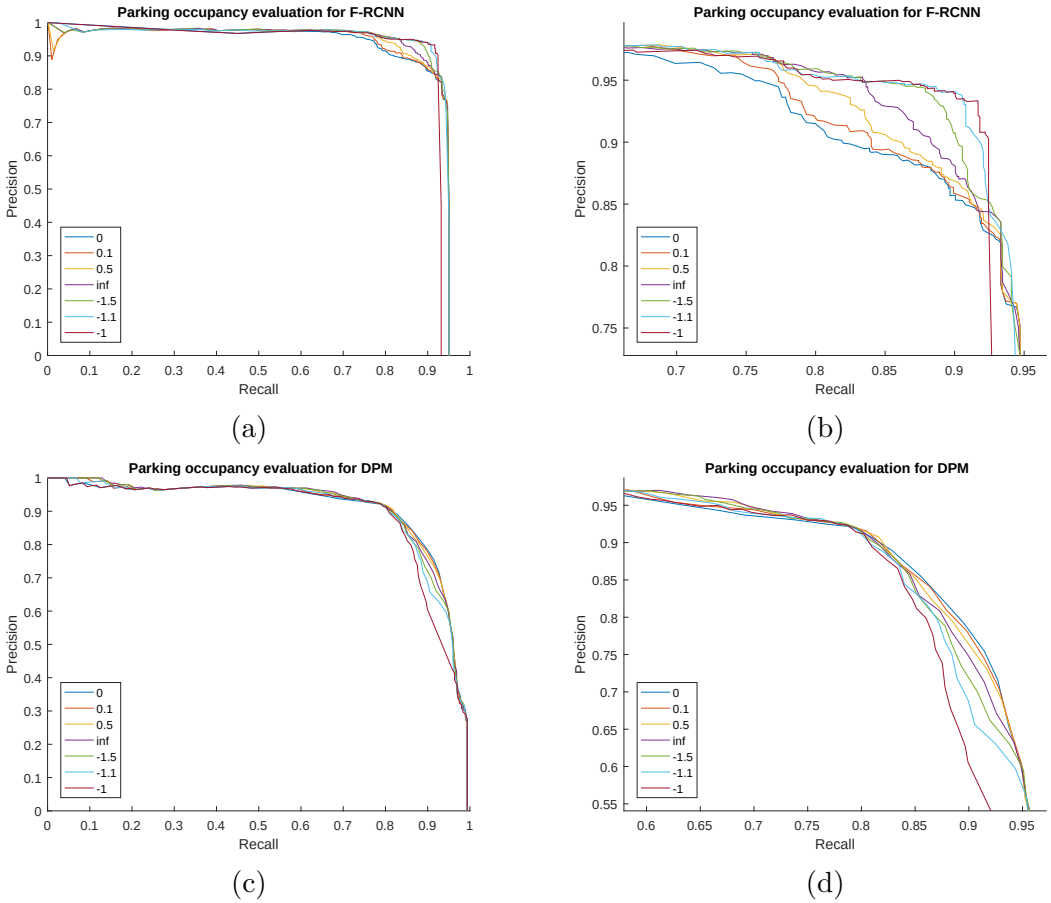


Fig. 5.11. Multi-camera parking occupancy evaluation for the two object detection trained models: full curve (left) and zoom of the equal error rate area (right).

k	0	0.1	0.5	∞	-1.5	-1.1	-1	**Optimal k**
F-RCNN	0.910	0.909	0.912	0.916	0.918	**0.919**	0.905	-1.1
DPM	0.910	**0.913**	**0.913**	0.912	0.910	0.905	0.894	0.5
Det. GT	**0.991**	0.982	0.982	0.982	0.982	0.983	0.980	0

Table 5.3: Multi-camera parking occupancy evaluation: Area Under the Curve for the two object detection trained models and for the ideal detection (detection ground truth), considering different k parameter for the normalized sigmoid functions. The best results obtained for each algorithm are shown in bold.

		k						
Alg.	**Dens.**	0	0.1	0.5	∞	-1.5	-1.1	-1
F-RCNN	High	0.960	0.961	0.962	0,.962	0.961	0.958	0.956
	Med.	**0.965**	**0.969**	**0.971**	**0.972**	**0.972**	**0.969**	**0.963**
	Low	0.863	0.867	0.870	0.871	0.872	0.869	0.860
DPM	High	0.918	0.919	0.919	0.920	0.920	0.919	0.892
	Med.	0.923	**0.924**	**0.924**	**0.925**	**0.926**	**0.928**	0.914
	Low	**0.930**	0.916	0.917	0.920	0.920	0.917	**0.926**

Table 5.4: Parking occupation density evaluation: Area Under the Curve for the two object detection trained models divided in three occupation density categories: low (1-12 vehicles), medium (13-24 vehicles) and high (25-36 vehicles).

		k						
Alg.	**Dens.**	0	0.1	0.5	∞	-1.5	-1.1	-1
F-RCNN	Day.	0.909	0.912	**0.913**	0.912	0.909	0.904	0.893
	Night.	**0.939**	0.935	0.927	0.924	0.930	0.931	0.928
	Clear	0.909	0.912	**0.913**	**0.913**	0.910	0.904	0.894
	Rainy	**0.917**	0.910	0.903	0.897	0.898	0.901	0.897
DPM	Day.	0.905	0.905	0.908	0.912	0.914	**0.915**	0.900
	Night.	**0.993**	**0.993**	**0.993**	**0.993**	**0.993**	**0.993**	0.972
	Clear	0.905	0.905	0.908	0.913	0.914	**0.915**	0.899
	Rainy	0.967	0.968	0.967	0.964	0.967	0.971	**0.973**

Table 5.5: Parking weather evaluation: Area Under the Curve for the two object detection trained models divided in four weather categories: daytime/nighttime and clear/rainy.

this evaluation, the weight of each camera would be learned from the k parameter and detection threshold with the best overall score. The computational and time cost of this evaluation is reduced and it does not require previous knowledge of the distance, reliability or quality of each camera for the person who is responsible for deploying and adapting the system.

Parking occupation density evaluation To verify that the proposed complete system is robust to occlusions, an additional study is added, classifying the frames in three occupancy density categories: low (1-12 vehicles), medium (13-24 vehicles) and high (25-36 vehicles). The results of this study are shown in Table 5.4. In spite of the need to divide a small number of frames (100) into three categories, which results in low resolution curves, the results show that the system performs correctly in occlusions situations. The results obtained by disaggregating the dataset are close to the mean except for Faster-RCNN low density occupancy, suffering a fall of performance of about 10% due to in scenarios with low vehicle density, a misclassification of a vehicle penalizes doubly (false positive and false negative) when occupying the square adjacent to the one it actually occupies. It should be noted that this system is especially designed for high

density scenarios, where it is most useful to route vehicles to places where there are available spots.

Parking weather evaluation Following the same procedure, a study of the system performance for different types of weather is added, classifying the frames in four weather categories: daytime/nighttime and clear/rainy. The results of this study are shown in Table 5.5. The scores obtained for the system in the nighttime frames are better than for the complete image set, due to during the night there are less reflections, which facilitates the operation of the detector. With respect to rainy frames, for the DPM detector the results get worse (between 0.003 and 0.015) for k = [0.1, 0.5, ∞, -1.5, -1.1] and improve (between 0.003 and 0.007) for k = [0, -1]. For the faster-RCNN detector, the results improve the overall performance between and 0.052 and 0.058 for all k. For daytime and clear frames sets, the behavior of the system is practically identical to the general behavior, as these categories contain most of the frames considered in the synchronized category. These small performance variations do not affect the system operation so, as indicated above, the system works for different types of lighting and weather conditions.

Optimal parameter k The process of the optimal k parameter learning for each algorithm consists of evaluating the range of values of the parameter, selecting the value that best adapts to the characteristics of the detection algorithm. After performing the experiments, considering the different possible values of k parameter, an optimal parameter has been obtained for each of the algorithms considered, as indicated in Table 5.3. For the Faster-RCNN algorithm, the best sigmoid is obtained with the parameter $k = -1.1$. This is due to the most useful information is generated in the spots closest to each camera, and for this reason this parameter generates the best score after the dataset evaluation. In the case of the DPM algorithm, the best sigmoid is obtained with the parameter $k = 0.5$. In this case, the combination whose experimental score is better is obtained by maintaining the detections of medium distance with a greater weight than that considered for the Faster-RCNN algorithm. The optimal examples of sigmoids, and other examples, are shown in Figure 5.7. An optimum value is also obtained experimentally with k = 0 for the case in which the detections were ideal. This result is consistent as in the case of ideal detections, all detections have the same confidence and are, therefore, weighted with a constant value (flat sigmoid).

5.5 Conclusions

This chapter presents a multi-camera system for the management of vacant parking places by means of vehicle detection and their corresponding mapping into the parking spaces of a parking lot. The system has been designed so that existing parking lot security cameras can be used for the proposed system after a simple configuration, without the need for a complete new camera

deployment. The designed system faces more complicated scenarios than the ones tackled in the state of the art: almost total occlusions and climatic changes (cloudy scenarios, rain, snow ...), that limits/reduces their performance. In this scenario with such a variable background it is not possible to carry out a precise background extraction, nor it is possible to label and define the region of each place as some parked vehicles completely occlude some of the spots behind them. In addition, the consideration of a multi-camera scenario, which, as far as we know, has not been reported before for this type of systems, is added.

There are multiple future work lines to improve the proposed system. With respect to the combination, we have chosen a simple technique using normalized sigmoid functions, therefore different functions could be studied in order to optimize the combination or fusion of the different information sources. Also a new dataset with more cameras and with different spatial configurations could be recorded to see the behavior of the system in those situations. A tracker can be added to the sequence detection to combine the information extracted during the sequence frames providing temporary continuity to the vehicle detections. Apart from this, current lines of future work for object detection can be applied here, since the detector is the first stage of the system.

Part IV

Detection Improvements in Multi-camera Scenarios

Chapter 6

Improving multi-camera people detection using contextual information

6.1 Introduction[1]

This Chapter combines information obtained from different cameras in order to enhance people detection algorithms. Using multiple cameras and information from the recorded scenario, called contextual information (distances between detected objects and cameras, position of the cameras, etc.), the detection performance is improved taking advantage of the results of the other cameras, transferring information from one camera to another, and then combining it. A cylinder is considered to approximate the person position and volume in order to place the person in the common plane and to transfer the position of the detection bounding boxes to other camera viewpoint. The proposed system has been evaluated on three different datasets obtaining improvements in the performance of the detection results, both for detections with different appearances and aspect ratios (as for example wheelchair users), as well as for scenarios with different camera placements and settings. The technique has also been adapted to vehicles detection, obtaining significant improvements in detection performance.

The structure of the following sections is as follows. Some related works are presented in Section 6.2, and Section 6.3 describes the proposed technique for the information transfer and combination between cameras. Section 6.4 presents the evaluation framework. Experimental results are discussed in Section 6.5. Finally, Section 6.6 summarizes the conclusions and future work.

[1]This chapter is an adapted version of the publications [Martín-Nieto et al., 2018c]

6.2 State of the art

People detection techniques can be divided into three stages [García-Martín and Martínez, 2015b]: firstly, a person model is designed which defines the characteristics that the detected objects must fulfill to be considered people; secondly, an object extraction process is performed, which will find the candidates to be classified; finally, the classification consists of the comparison of the objects detected in the sequence with the model generated in the first step. In this step, a decision is made on the objects and it is decided whether the objects are classified as persons or not. Depending on the application, the decision can be person vs. non person, or a probability value of being a person.

The information provided by a single camera is limited, so in order to monitor a wide area or to obtain more information from the different viewpoints of a region of interest, it is necessary to use more than one camera. For this reason, the use of several cameras is a common way of developing applications, since it is also useful for solving occlusions in scenarios with high density of people and for 3D applications.

The key technologies when using multi-camera are, as explained in [Wang, 2013]: calibration of all cameras in a single coordinated system; knowledge of the topology of the camera network to get information on how they are related to each other; identification of objects in several cameras to determine if the observed objects are the same or different; tracking objects across all cameras; automatic recognition of abnormal actions or activities.

The use of a multi-camera environment in scenarios with possible occlusions, usually improves the detection performance with respect to the use of the cameras independently. A method is proposed in [Santos and Morimoto, 2008] to perform detection and tracking of people in multi-camera environments where there are occlusions. This method is based on the methodology proposed in [Kim and Davis, 2006] and is based on using the information of each of the cameras from the scenario, merging it into a common plane (the ground plane) obtained by homographies. The individual information that is combined in the common plane is previously obtained by subtracting the background. Then, the object detections are performed in the common plane and, afterwards, the correspondence between cameras and objects is made. In this way, using cameras with different locations, the problem of occlusion is generally solved. The main limitation is that the individuals have to appear initially isolated. In [Santos and Morimoto, 2011], an improvement of the previous method [Kim and Davis, 2006] to eliminate false positives is proposed. The algorithm that performs this process compares the views of all the cameras for each one of the detected objects, thus reducing the false detections, applying multiple view perspective geometry of people presence on the ground plane. It is also interesting to consider [Black et al., 2002], where a method that uses a Kalman filter to obtain 3D information from the 2D information is presented.

Unlike the previous approaches, we transfer the detections from one camera to another

instead of just projecting all the detections to the common plane. In this way, the object information is not reduced to a simple coordinate, allowing to transfer more information (volume, height, aspect ratio) and to process the information for each camera viewpoint. Instead of processing the information only in the common plane, the detections from each camera are previously combined with the information from the other ones, which can also be used to further combine the information in the common plane with the additional advantage of having previously improved and corrected the information of each camera.

In the state of the art, the information of the detections is usually projected in the ground plane at point level (one point per detection) or at mask level (masks are projected and the intersections indicate the position of the detected person). The work presented in this chapter considers the common plane to obtain the different camera views information, and it allows to transfer (and combine) object detections from each camera to the other ones.

6.3 Proposed technique

6.3.1 Cylinder estimation and information transfer

A cylinder is considered to approximate the person position and volume in order to transfer the position of the detection bounding boxes from one camera to another, maintaining the volume that occupies a person, instead of using only the projected plane generated from the detected bounding box. The consideration of the representation of people as cylinders has been used previously in the state of the art [Kilambi et al., 2008], but as a method for people counting (estimation) from a single camera perspective. The objective of the developed technique is to transfer the bounding boxes of the detections from one camera to the viewpoint of another camera. As the projections on the common plane of the detected bounding boxes do not correspond spatially with the position and volume of the detected object, the transfer between cameras must be corrected. Figure 6.1(c) shows a case in which the transferred bounding box (blue) does not fit the person when changing the point of view, so it is necessary to process it to obtain the correct box (red). Here we describe the method applied to each bounding box detected by the camera whose information is transferred.

1. Firstly, the base (bottom) segment of the detection bounding box is projected to the common plane. This plane can be obtained using homographic techniques, or from the intrinsic and extrinsic parameters of the cameras. We use the base segment as it is in the common plane, which allows to accurately transfer it. Figure 6.1(a) shows two bounding boxes which will be transferred.

2. Using the projected segment in the common plane, a circumference is defined so that the projected segment forms one of the sides of a square inscribed therein. In Figure 6.1(b),

the projected segment is represented with the blue line, the square is represented with discontinuous blue line, and the circumference is represented with a (green) circle.

3. To define the bounding box base segment which will be transferred to the other camera, the inscribed square (blue) is rotated (represented with discontinuous red line in Figure 6.1(b)) with an angle such that the closest side is perpendicular to the line connecting the new camera with the center of the circumference (green cross in Figure 6.1(b)). This side (red line) corresponds to the projection of the transferred bounding box base segment.

4. The height of the cylinder is estimated using proportionality, taking into account the object original height and the cameras distances to the object.

5. Finally, this generated cylinder is transferred to the point of view of the new camera, again using an homography (inverse matrix) or from the intrinsic and extrinsic parameters of the new camera. An example of the resulting cylinders is shown in Figure 6.1(c). Another example of the resulting cylinders for people with different appearance (standing person and wheelchair user) is shown in Figure 6.2.

6.3.2 Detections combination

It is common for an object to be detected in several cameras, so it is necessary to add an information combination stage. The multiple matching bounding boxes of the same person are simplified into a single one. The measures used for this association are the same as the measures used in the evaluation to decide if two bounding boxes correspond to the same object (see subsection 2.5.1). Two detection bounding boxes are considered to correspond with the same person if $rd \leq 0.5$ (relative distance between bounding boxes, corresponding to a deviation up to 25% of the true object size) and cover and overlap between bounding boxes are both above 50%. More details of these commonly used metrics are available in [Leibe et al., 2005]. When two bounding boxes are associated as belonging to the same object, the one with greater confidence is maintained, and the other one is deleted to eliminate redundancy.

Figure 6.3(c) shows examples of bounding boxes of the ground truth (red), the own camera detections (green) and the transferred detections by the other camera (blue). In this case, each camera separately is capable of detecting only two people (6.3(a) and 6.3(b)). For the person in the middle of the image, the bounding boxes of the two cameras (blue and green) are combined into a single one, resulting in a final complete detection containing three bounding boxes, one

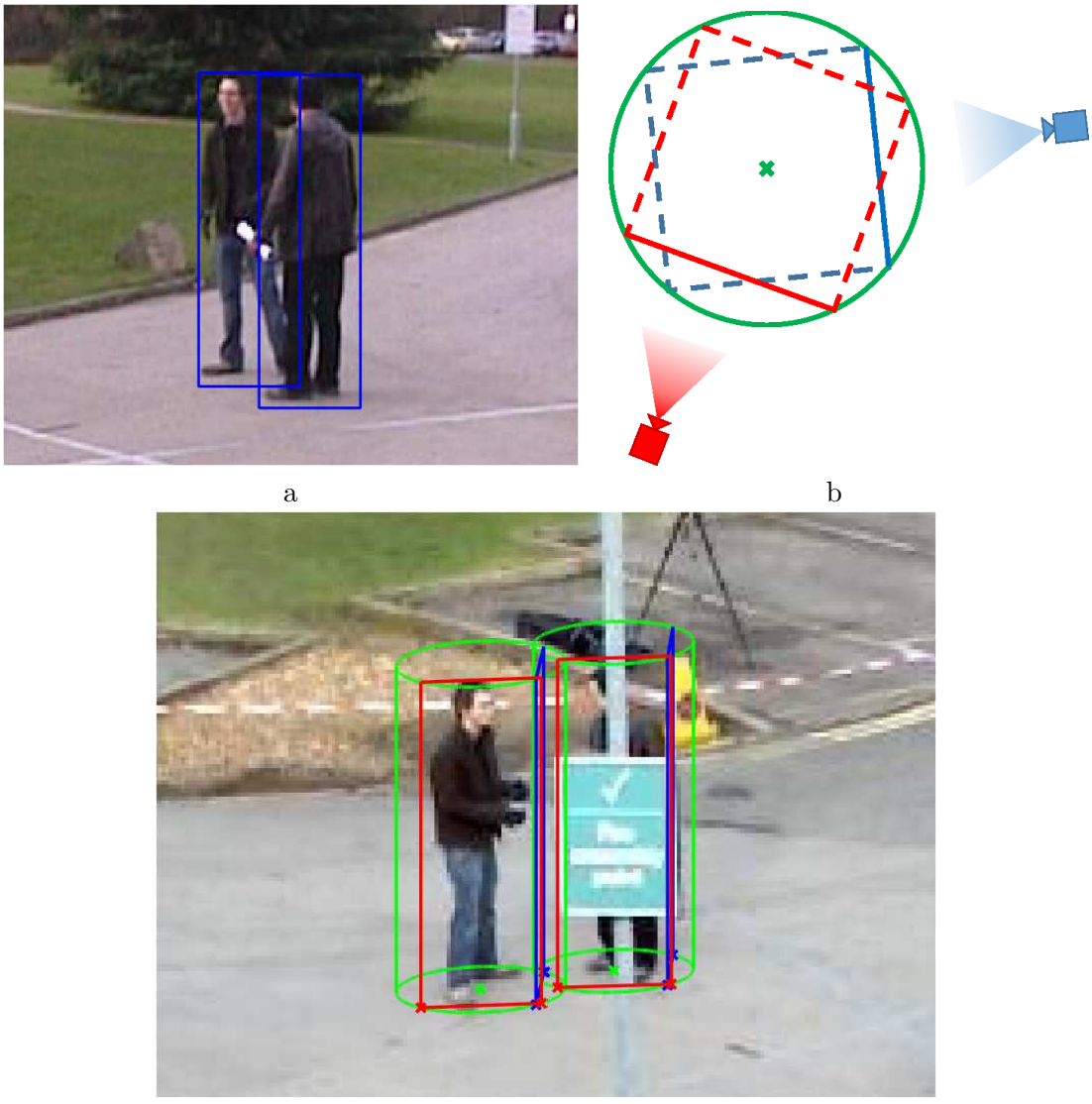


Fig. 6.1. Overview of the proposed technique process. (a) shows two detection bounding box examples, (b) schematizes the geometric process, and (c) contains a representation of the resulting cylinders (green), the original bounding box (blue, very tilted due to the angle between the cameras viewpoints), and the resulting bounding boxes (red).

Fig. 6.2. Example of cylinders for people with different aspect ratio

for each person, each one being very similar to those annotated manually in the ground truth (red). A complete result is obtained, eliminating redundancy between cameras detections.

6.4 Evaluation framework

The method proposed in this chapter works at bounding box level, so it can be applied to any detector whose outcome is composed by a list of bounding boxes.

For the evaluation of the method, we have decided to use the algorithm known as DPM (Deformable Part Models) [Felzenszwalb et al., 2010b]. The reason to choose this algorithm is that we had a previously trained wheelchair user model, which allowed to evaluate the results on a dataset with two completely different people appearances (WUds which contains standing people and wheelchair users, see subsection 2.3.1). Given an object, it may be divided into parts having common properties, while the deformable component may be characterized by the connection between pairs of neighboring parts. In case the different appearance models were available for another detection algorithm, it could be used in the same way as the selected one.

In order to further evaluate in a more realistic scenario, PETS 2009 Benchmark sequences have been used (see Subsection 2.2.1). This dataset contains outdoor sequences from a typical surveillance setup. The cameras are calibrated using the Tsai calibration [Tsai, 1986] and the calibration files are included in the dataset. We also evaluate this technique with the EPFL-RLC dataset (see Subsection 2.2.2). It was recorded in the EPFL Rolex Learning Center using three static HD cameras. The complete ground truth was not available so we manually annotated the bounding boxes of the detections for the first 2000 frames of camera 1. We make this ground

a b

c

Fig. 6.3. Detection combination example. (a) shows the own camera detections, (b) shows the transferred detections, and (c) represents the own (green), the transferred (blue) and the ground truth (red).

truth publicly available upon request.

6.5 Experiments and results

6.5.1 Camera viewpoint results

This subsection presents the results of the proposed technique for information transfer and combination between cameras at camera viewpoint level.

Firstly, in order to verify that the proposed technique works correctly on detections with different appearance and aspect ratio, the evaluation on the WUds dataset is made. This is performed by detecting people with a standing person model and with a wheelchair user model, transferring the information to the other camera and evaluating, in both cameras, the own camera detections, the detection transferred from the other camera and the information combination from both cameras. Figure 6.4 shows the precision-recall curves and Table 6.1 (WUds column) shows the AUC value for each curve. The detections obtained by the evaluated cameras are better than the detections transferred from the other camera and the information combination improves the results of each camera separately. This scenario is relatively simple, since there are not too many people in the scene and, therefore, the results are relatively good in all the three cases, but it shows that the proposed technique works for detection of objects with different aspect ratio and an improvement is achieved by the proposed information combination.

The proposed technique is evaluated also on the PETS2009 sequences. As discussed in the dataset subsection (2.2.1), in this case the only available ground truth is from view 1, so all evaluations are performed with respect to this camera. Figure 6.5 shows the precision-recall curves and Table 6.1 (PETS2009 columns) shows the AUC value for each curve. In this case, the improvement obtained is higher than in the previous dataset. The camera facing the evaluated viewpoint is view 8. It is closer to the monitored zone than view 1 (that is why it performs better than the reference camera 1) and the combination of both improves the final result score. The best result is obtained by combining view 6 with view 1. View 6 has a similar orientation to view 1 and, being closer to the scene, it gets the best individual score. Despite having the same orientation, the combination of these two cameras obtains the best AUC of all curves as they record the same area from different distances. This is due to the small viewpoint change, which limits the possible errors when translating the detections. View 5 and view 7 are very similar to each other, as they correspond to the side views, one on each side with respect to view 1. Therefore, their results are very similar to each other, being slightly better the behavior of view 5 and its subsequent combination with view 1. It should be noted that in all camera combinations, the results obtained are better than those of any camera separately.

Finally, for the EPFL-RLC dataset, the scenario has similar characteristics to the WUds scenario (reduced distances, indoor, etc.), with the difference of a greater people density and a

		WUds	PETS2009				EPFL-RLC	
Own view		0.73	0.62 (Cam1)				0.44 (Cam0)	
Transf. view		0.72	Cam5	Cam6	Cam7	Cam8	Cam1	Cam2
			0.46	0.71	0.45	0.70	0.43	0.29
Combined		**0.77**	0.78	**0.86**	0.72	0.80	**0.61**	0.49
Gain (Δ%)	v. own	4.7	20.1	28.0	14.0	22.1	38.8	12.1
	v. transf.	7.0	40.3	17.9	37.8	11.8	41.2	73.5

Table 6.1: AUC Results for camera viewpoint evaluation of WUds, PETS2009 and EPFL-RLC.

		WUds	PETS2009				EPFL-RLC	
Own view		0.76	0.60 (Cam1)				0.37 (Cam0)	
Transf. view		0.76	Cam5	Cam6	Cam7	Cam8	Cam1	Cam2
			0.44	0.63	0.32	0.60	0.43	0.29
Combined		**0.83**	0,72	**0,76**	0,63	0,68	**0.65**	0.47
Gain (Δ%)	v. own	9.8	21.0	26.4	6.0	13.6	74.9	28.3
	v. transf.	9.7	64.5	20.9	96.5	14.0	50.9	62.9

Table 6.2: AUC Results for ground plane evaluation of PETS2009 and WUds.

larger number of occlusions. This greater complexity of the scenario implies greater difficulty for the proper system operation and, therefore, the values of AUC are lower than those obtained in the first dataset. Thanks to the multi camera combination, it is possible to solve these occlusions, which allows the improvement of the scores.

6.5.2 Ground plane results

This subsection presents the results of the proposed technique for information transfer and combination between cameras at ground plane level.

With respect to the evaluation of the ground plane, the results obtained are similar to those of the camera viewpoint evaluation, but with some differences that are discussed below. For the Wheelchair Users dataset, the obtained results are slightly better than those obtained for the camera viewpoint, due to the existence of a reduced scenario (room) and relatively short distances. For this dataset, the obtained error when projecting the detections into the ground is reduced, obtaining a better association between GT and detection blobs, thanks to adding a new spatial dimension that completes and improves the information obtained from the camera viewpoint plane. In the case of the PETS2009 dataset, as it is recorded in an open space, and due to the greater distances between cameras, the score obtained for the evaluation of the ground plane is lower than the camera viewpoint scores, as accuracy errors increase when projecting in a larger ground plane. Despite this, the multi-camera combination keeps improving the cameras performance separately, and the comparative behavior of each obtained PR curve is similar to

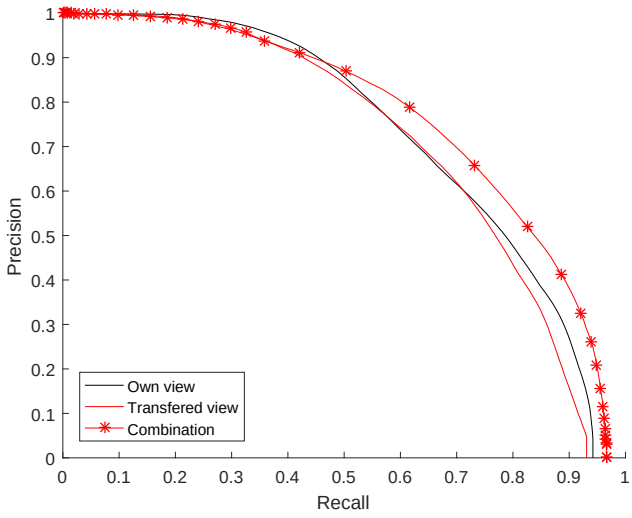


Fig. 6.4. AUC curves for the WUds camera viewpoint evaluation.

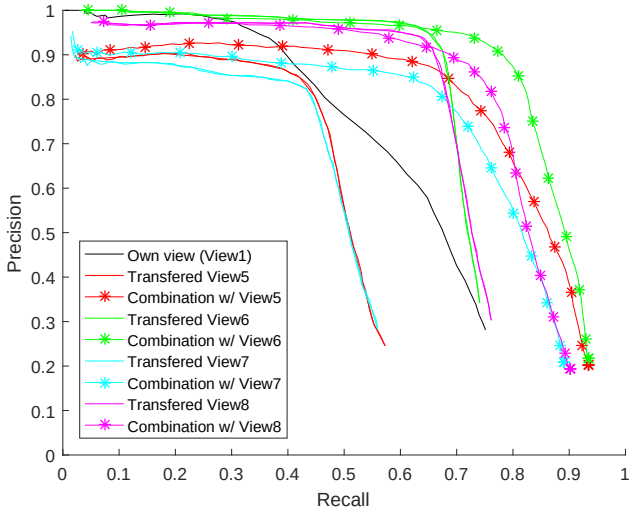


Fig. 6.5. AUC curves for the PETS2009 camera viewpoint evaluation.

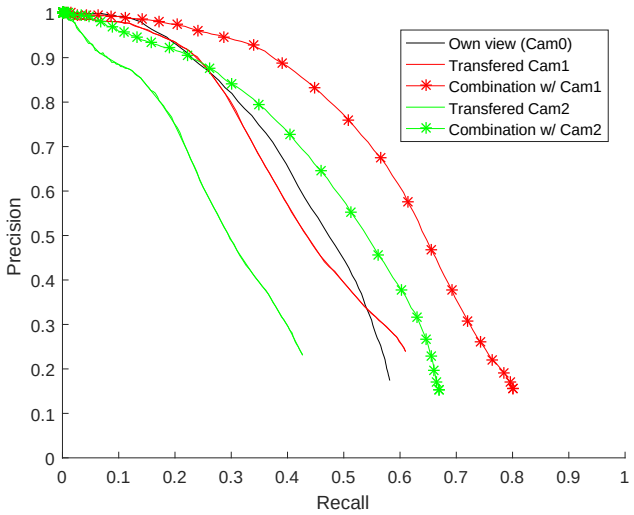


Fig. 6.6. AUC curves for the EPFL-RLC camera viewpoint evaluation.

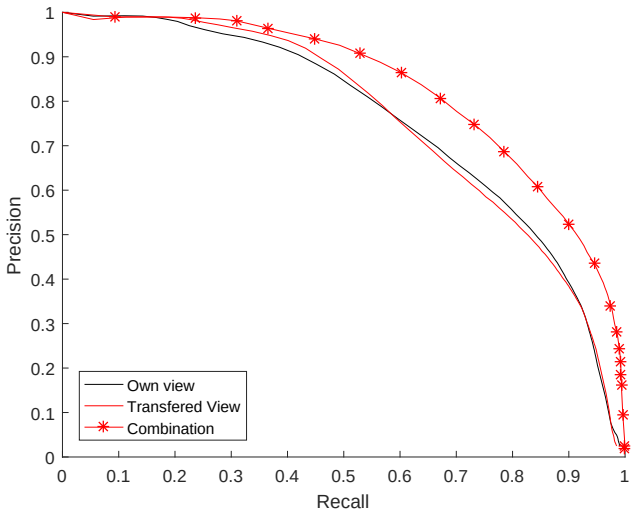


Fig. 6.7. AUC curves for the WUds ground plane evaluation.

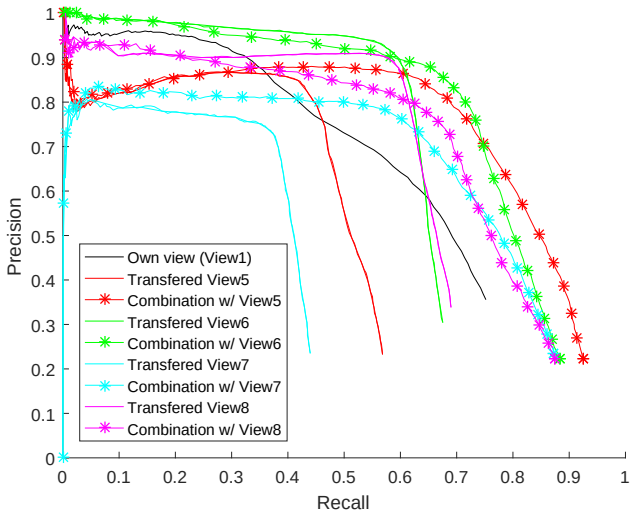


Fig. 6.8. AUC curves for the PETS2009 ground plane evaluation.

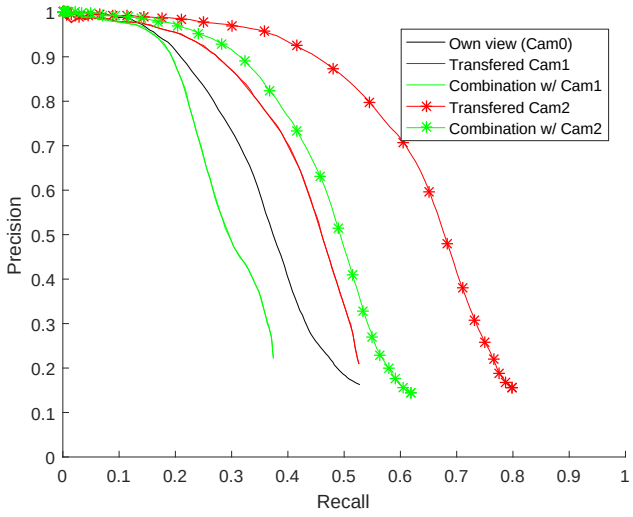


Fig. 6.9. AUC curves for the EPFL-RLC ground plane evaluation.

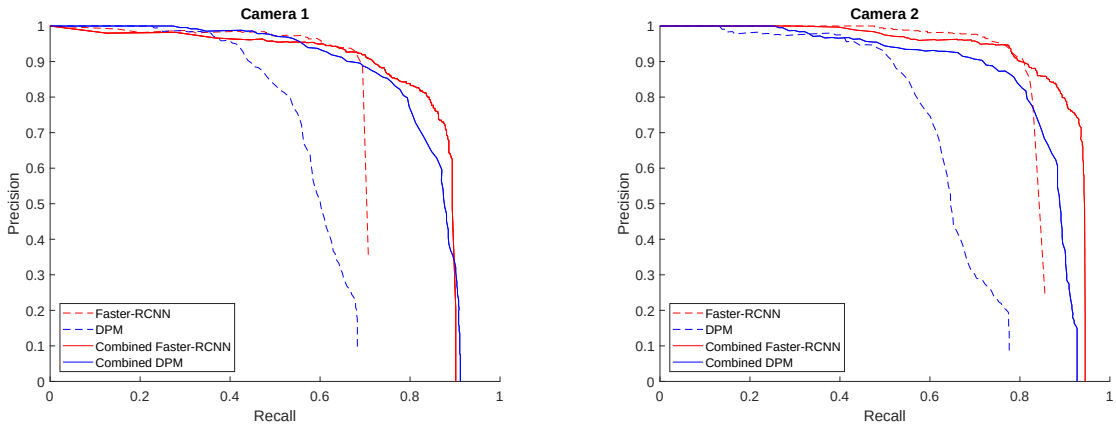


Fig. 6.10. AUC curves for the PLds camera viewpoint evaluation.

	Camera 1		Camera 2	
	DPM	Faster-RCNN	DPM	Faster-RCNN
Own view	0.586	0.687	0.6426	0.830
Combined	0.831	0.839	0.8410	0.904
Gain (Δ%)	41,8	22,2	30,9	8,8

Table 6.3: AUC Results for camera viewpoint evaluation of PLds.

the previous evaluation (camera viewpoint evaluation). Finally, the EPFL-RLC PR curves have a similar appearance for both evaluations, with the difference that the score obtained from the evaluation of the ground plane for camera 1 is better than the score obtained from the ground evaluation of camera 0, in which it is evaluated.

6.5.3 Vehicle detection results

This subsection shows the result of applying the ideas in this chapter to vehicle detection. As vehicles have a very different appearance than people, the techniques here applied are modified to maintain the technique functionality. For the information transference, the initial part of the procedure used in Chapter 5 (see Figure 5.1) but at the output of the perspective correction block the process is carried backwards with the parameters of the other camera. The final result of this process is the same as that described in this chapter, detections transferred from one camera to another. For this evaluation we consider the PLds sequences (see section 2.3.2). After evaluating in the same way as the people detections evaluation, the results obtained are shown in Figure 6.10 and Table 6.3.

The results show that the applied technique improves the vehicles detection, both for the detector with the worst performance (DPM) and for the detector with the highest performance (Faster-RCNN). The utility of combining again the detections to evaluate at spots level, as

performed in Chapter 5, is not clear, as combining the same information twice can lead to redundant errors.

6.6 Conclusions

This chapter presents a multi-camera system to enhance people detection algorithms. The system adds contextual information of the scene to a people detector, resulting in an improvement of the performance of each camera independently. The results obtained after the evaluation of the system (camera viewpoint evaluation and ground plane evaluation) confirm the initial hypothesis, obtaining improvements in the detections performance by combining the cameras information in the three evaluated scenarios. The proposed method works for different people aspect ratio (standing, sitting, etc.) and for any orientation between the different cameras thanks to the proposed volumetric assumption (cylinder person approximation). The technique has also been adapted to vehicles detection, obtaining significant improvements in detection performance. The cylinder estimation and information transfer stages are completely parallelizable for each camera, so the increase in computational cost with respect to a standard system is the detections combination stage.

As future work, it is proposed to use and combine other detectors (e.g. the ones cited in subsection 6.4) to try to improve the system performance and to develop other more elaborated information combination methods, for example by weighing the contribution of each camera to the combined detection as a function of distance between the camera and the object/person, since the closest detections to a camera are usually more accurate and have a greater confidence. In addition, a greater number of cameras can be combined simultaneously, controlling and eliminating the errors introduced by the detections of each camera, using, for example, a majority voting system to decide which detections are kept and which are discarded. Alternatively, the presented technique can be applied to other objects instead of only people, as for example vehicles. In this case, instead of a cylinder, the occupied volume could be considered as a cuboid, and the steps of the process should be adapted to this new object.

Chapter 7

Enhancing multi-camera people detection by stand-alone automatic parametrization using detection transfer and self-correlation maximization

7.1 Introduction[1]

By employing multiple cameras for video object detection, the available viewpoints provide additional information that may allow to overcome the limitations of detectors applied to single camera views. However, determining the confidence of the information generated for each viewpoint remains a challenging problem. In this chapter, we propose a framework to automatically adapt and improve any detector in multi-camera scenarios, during runtime detection, where people are observed from various viewpoints. Unlike generic approaches fixing confidence thresholds, this framework adapts the detector's threshold for each frame and camera. The framework considers a detector applied to each viewpoint with multiple thresholding hypotheses and then all results are transferred to a desired viewpoint. Later, correlations are computed for each pair of transferred results, which determines an optimal threshold for each pair of cameras. Finally, the pair-wise selected thresholds are combined by weighted voting to obtain the best adapted threshold for each individual camera. We consider generic threshold-based detectors, pre-trained on standard datasets, and a cylindrical model, making this proposal applicable to most state of the art people detectors. The experimental results demonstrate that the proposed

[1]This chapter is an adapted version of the publications [Martín-Nieto et al., 2018b]

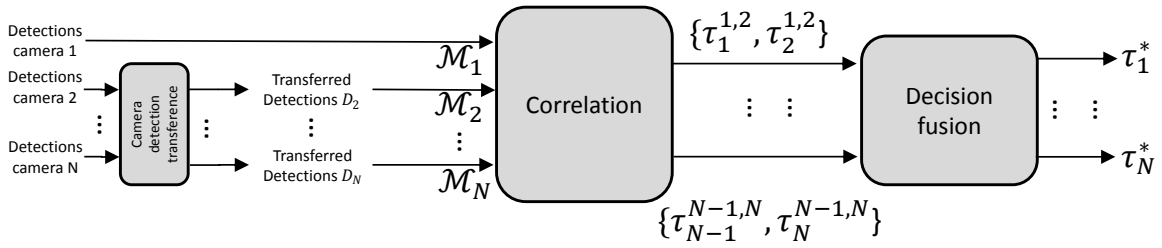


Fig. 7.1. Framework overview.

framework improves the performance of state-of-the-art detectors whose optimal parameters are determined during training time. The technique has also been adapted to vehicles detection, automatically obtaining the detection threshold, and significant improvements for the detection performance.

The structure of this chapter is: after this introduction, Section 7.2 presents an overview of the proposed technique. Section 7.3 describes the detection transference approach, which is a short description of the previous chapter. The correlation framework is described in Section 7.4. The evaluation and results of the approach are presented in Section 7.5. Finally, Section 7.6 contains conclusions and future work.

7.2 Framework overview

In this chapter we provide an integrated framework with optimal automatic parameterization, which improves "classic" cameras combination results in terms of direct transfer of bounding box. The different parts of the framework, which are described in more detail in their corresponding sections, are the following:

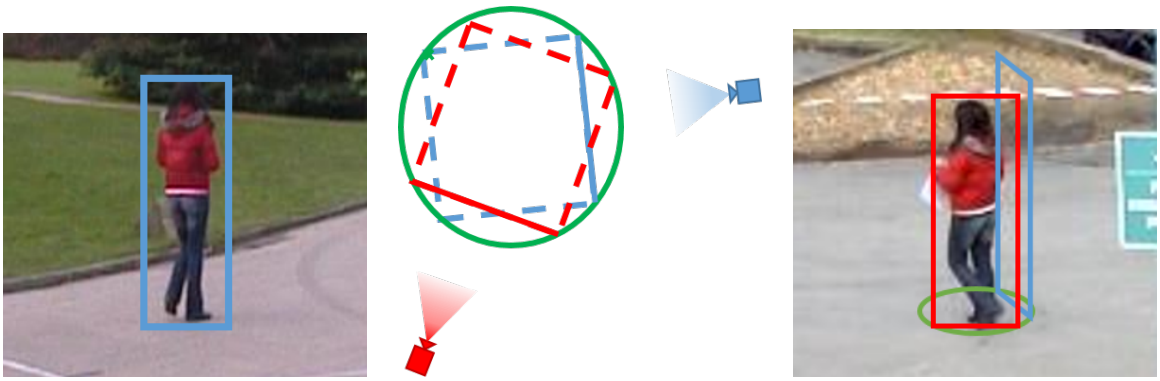
1. The detector is executed on the frames of all the cameras.

2. The detections of all cameras are transferred and homogenized to the desired viewpoint.

3. The homogenized detections from the previous stage are correlated frame by frame.

4. An optimal decision threshold is selected for each camera and frame, based on the correlation obtained in the previous step.

Figure 7.1 shows the complete framework with the steps described above.



(a) (b) (c)

Fig. 7.2. Detection transference example. (a) shows a detection bounding box example, (b) represents the geometric rotation, and (c) shows the original bounding box (blue), the resulting bounding box (red), and the rotation circle (green).

7.3 Detection transference between cameras

A cylinder is considered to approximate the person location in order to change the orientation of the detection bounding boxes from one camera to another. The consideration of the representation of people as cylinders was previously considered in the state of the art [Kilambi et al., 2008], but for people counting (number estimation) from a single camera perspective. Fig. 7.2(a) shows a case in which the transferred bounding box (blue) does not fit the person when changing the point of view Fig. 7.2(c) , so it is necessary to process it to obtain the correct bounding box (red). The transformation applied to each detection is: Using homographic techniques, or from the intrinsic and extrinsic parameters of the cameras, the bottom part of the bounding box of the detection bounding box is projected to the common plane. A circumference is defined so that the previously projected segment forms one of the sides of a square inscribed therein. The inscribed square is rotated to obtain a new square such that two of its sides are perpendicular to the point of view of the new camera. Finally, this generated bounding box is transferred to the point of view of the new camera, using the inverse process applied in step 1 Fig. 7.2(a) shows a bounding box which will be transferred. In Fig. 7.2(b), the projected bounding box base is represented with the blue line, and the cylinder base is represented with a (green) circle. The red line corresponds to the projection of the transferred bounding box base and belongs to the rotated (red) square. An example of the resulting cylinders is shown in Fig. 7.2(c).

7.4 Correlation framework

We apply a framework to improve the performance at runtime detection by adapting the detector configuration (see Figure 7.1). This proposal is based on the maximization of mutual information

strategy where classifiers are combined assuming that their errors are complementary [García-Martín and SanMiguel, 2017]. In our case, the detection model, executed in the different cameras, has been trained using the same content set. The incorrect detections will be different for each camera, so the correlation will reinforce the correct detections common to all cameras and penalize the isolated errors of each camera.

We start from a set of N camera frames, whose detections are transferred to a single main camera ${D_n}_{n=1}^{N}$. Each camera detection D_n obtains a confidence map, M_n, representing the likelihood for each spatial location in the image. Then, detection candidates are obtained by thresholding this map. The transferred camera detections are compared to obtain a set of pairwise correlation scores. First, the decision space of each camera output is explored by applying multiple thresholds. Then, these multiple outputs are correlated for each pair of camera detections (D_n and D_m) to obtain a correlation map which measures the output similarity. Finally, the configuration with the highest similarity allows to select the best detection threshold for each camera output ($\tau_n^{n,m}$ and $\tau_m^{n,m}$, respectively). Up to this point, we have hypothesis obtained for each compared pair of detections (D_n and D_m) which are combined to obtain a final configuration for each camera detections ($\tau_1^*...\tau_N^*$). Such hypothesis combination is performed as a traditional mixture of experts via weighted voting in the decision fusion block.

The correlation is only coherent to be carried out in the common area of the cameras, since otherwise disjoint sets would be correlated and the process would not be useful. To locate the common area, the ground plane of each camera is transferred to the desired point of view. An example of this process is shown in Figure 7.3, in which the plane of each camera is represented with a different colour, and the area common to all the cameras has been darkened for facilitating its localization.

7.5 Experiments and results

The objective of this section is to evaluate the presented framework in order to numerically demonstrate the detection improvements. Two datasets have been considered for the evaluation, which contain overlapping multi-camera environments. PETS 2009 dataset (see Subsection 2.2.1) presents outdoor sequences from a typical surveillance setup. We consider the available ground truth from [Milan et al., 2014] and sequences S2-L1 and S3MF1 which are the only ones that contain available and synchronized frames for cameras 1, 5, 6, 7 and 8 (five cameras in total). The EPFL-RLC dataset (see Subsction 2.2.2) was recorded in the EPFL Rolex Learning Center using three static HD cameras. The complete ground truth was not available so we manually annotated the bounding boxes of the detections for the first 2000 frames of camera 1. We make this ground truth publicly available upon request. Both datasets are calibrated using the Tsai calibration [Tsai, 1986] and the calibration files are included in the respective websites

Fig. 7.3. PETS2009 View plane of each camera and common area for all cameras.

of the authors. The detection performance is evaluated by Precision, Recall and FScore metrics for each frame (see Subsection 2.5.1). We consider the mean FScore for all sequence frames as the final performance value. With respect to the detection algorithms, we consider four people detectors with publicly available implementations: DPM [Felzenszwalb et al., 2010b] (Inria model), Faster R-CNN [Ren et al., 2015] (VGG model), ACF [Dollar et al., 2014] (INRIA model) and YOLO9000 [Redmon and Farhadi, 2017].

The results obtained after evaluating the four detectors on the sequences of the two datasets are presented in Table 1. Best fixed threshold columns show the FScore obtained using the threshold selected using the best a priori knowledge (fixed during training time) for each viewpoint. The improvement obtained by applying the transfer of detections between cameras and the correlation framework is greater for detectors with worse performance (in this case DPM, followed by ACF), as expected because the improvement margin is greater. For better performance detectors (Faster R-CNN, YOLO9000) an improvemement of the results is also achieved for all cases. The technique has also been adapted to vehicles detection, obtaining significant improvements in detection performance.

In addition to these results, the great advantage obtained by eliminating the selection of the confidence detection threshold, which is not reflected in the table or in the results, should be taken into account.

		DPM			Faster R-CNN		
Sequence	#Cameras	Best fixed	Ours	%Δ	Best fixed	Ours	%Δ
PETS S2-L1	5	0,35	0,45	**29,0**	0,69	0,74	**7,6**
PETS S3MF1	5	0,38	0,50	**31,8**	0,65	0,66	**2,6**
EPFL-RLC	3	0,32	0,49	**54,6**	0,65	0,73	**12,9**

		ACF			YOLO9000		
Sequence	#Cameras	Best fixed	Ours	%Δ	Best fixed	Ours	%Δ
PETS S2-L1	5	0,69	0,74	**6,5**	0,68	0,74	**9,9**
PETS S3MF1	5	0,49	0,61	**25,6**	0,60	0,65	**8,7**
EPFL-RLC	3	0,45	0,54	**20,4**	0,67	0,74	**11,2**

Table 7.1: F-score values obtained for the four detection algorithms and the three considered sequences.

		DPM			Faster R-CNN		
Sequence	#Cameras	Best fixed	Ours	%Δ	Best fixed	Ours	%Δ
PLds, Camera 1	2	0,402	0,490	**21,9**	0,765	0,830	**8,5**
PLds, Camera 2	2	0,298	0,507	**70,4**	0,638	0,795	**24,5**

Table 7.2: F-score values obtained for the vehicle detections in PLds.

7.5.1 Vehicle detection results

As performed in the previous chapter, vehicle detection transference between cameras can also be considered, applying the rest of the techniques described in this chapter. Starting from the detections transferred from one camera to another, using the PLds sequences (see section 2.3.2), we obtain the autoparameterization of the cameras, and a detection performance improvement. The results obtained are shown in Table 7.2 demonstrating that the techniques described in this chapter works for other detected objects different from people.

7.6 Conclusions

We present a framework to automatically parameterize people detectors during runtime. This proposal exploits the correlation among multiple camera detections transferred to a common camera to determine the best threshold for each camera. The proposed approach is capable of working over standard state of the art detector outputs (bounding boxes), so any kind of detector and object model can be considered. The cylinder model needs to be adapted as seen in the vehicle information transference, but despite this the method still obtains performance improvements when transferring information between cameras. This framework allows the automatic threshold parametrization without requiring any model (re-)training process and therefore is completely standalone. The computational cost is increased by a factor equal to the number

of cameras, but the processing time can be maintained as the processing of each camera and the correlation process are independent and parallelizable. For future work, more object detectors can be considered. Also multiple detectors and cameras could be combined simultaneously, in order to further improve the results.

Part V

Conclusions

Chapter 8

Achievements, conclusions and future work

8.1 Summary of achievements and main conclusions

This thesis has addressed different aspects related to object detection at three different levels: training and evaluation frameworks (Part II), detection approaches and applications (Part III) and detection improvements in multi-camera scenarios (Part IV).

First, with respect to the training and evaluation frameworks (Chapter 2), we have analyzed published previous work and existing datasets that met our requirements have been selected, which have been completed with the creation of two new datasets: one dataset considering people with different appearances (standing or using a wheelchair), and one dataset recording a real parking lot with samples of different illumination (day, night, sunrise with shadows) and weather (sunny, rainy) conditions. With respect to evaluation metrics, we have presented metrics commonly used for the evaluation of object detectors. First, the "classical" evaluation metrics have been formulated, and then we have adapted these metrics for, on the one hand considering a third dimension (depth) in the scenarios, and on the other hand, evaluating the capacity to detect occupied or empty parking spots. Finally, the detection algorithms used in the thesis have been presented and described.

Secondly, the motivation of Chapter 3 is the study of feasibility of training an object detection algorithm using a synthetic images dataset. In particular, the chosen object to train is a wheelchair user from empty wheelchairs images and standing people images. Three synthetic image datasets have been created in order to train three different models, evaluating which model is optimal and finally analyzing its feasibility by comparing it with a people detector for wheelchair users trained with real images. With the obtained results it can be concluded that the performance of the detectors is acceptable, although worse than the one obtained with the

original wheelchair people detector, trained with real images. This result was expected a priori, as the images that have been generated are synthetic and are different from the real recorded images, but despite this, a detector model has been obtained that is able to detect in an appropriate way. In exchange for this performance loss, a functional detector has been obtained without the need to record the real object (in this case, wheelchair users). This method can be useful in situations where it is not possible to compile or record a dataset of the desired object type, or obtain it is too expensive in terms of time or resources.

Thirdly, we treat the problem of different appearances for the same semantic object class detection (Chapter 4). Typical senior residences scenarios are an example of this problematic situation. In particular, our main objective is to detect both standing people and wheelchair users simultaneously. For this reason, an extension of people detection that allows to detect people with the need of using a wheelchair has been presented. We have trained two additional wheelchair users detectors models whose detections can be combined with the detections obtained using the traditional standing people detectors models, providing generality and complementary detection capacity. This approach can not only be applied to the case of wheelchairs but the ideas exposed here can be extrapolated to other scenarios where there are individuals with an appearance different from the standard, as Zimmer frames users or people using walking sticks.

Describing a different application for object detectors, Chapter 5 presents a multi-camera system for vehicles detection and their corresponding mapping into the parking spots of a parking lot. The system has been designed so that existing parking lot security cameras can be used for the proposed system after a simple configuration, without the need for a complete new camera deployment. The designed system faces more complicated scenarios than the ones tackled in the state of the art: almost total occlusions and climatic changes (cloudy scenarios, rain, snow ...), that limits/reduces their performance. In addition, the consideration of a multi-camera scenario, which, as far as we know, has not been reported before for this type of multi-camera systems, is added.

The first chapter of the Advanced Multi-Camera Techniques for Detection part (Chapter 6) presents a system to enhance object detection algorithms output, transferring and combining information obtained from different cameras. The system adds contextual information of the scene to a object detector, resulting in an improvement of the performance of each camera independently. The results obtained after the evaluation of the system (camera viewpoint evaluation and ground plane evaluation) confirm the initial hypothesis, obtaining improvements in the detections performance by combining the cameras information in the four evaluated scenarios. The proposed method works for different people aspect ratio (standing, sitting, etc.) and for any orientation between the different cameras thanks to the proposed volumetric assumption (cylinder person approximation). The cylinder estimation and information transfer stages are

completely parallelizable for each camera, so the increase in computational cost with respect to a standard system is the detections combination stage.

Continuing this last work described, a correlation framework to automatically parameterize object detectors during runtime has been added to the information transference between cameras (Chapter 7). This proposal exploits the correlation among multiple camera detections transferred to a common camera to determine the best threshold for each camera. The proposed approach is capable of working over standard state of the art detector outputs (bounding boxes), so any kind of detector and object model can be considered. This framework allows the automatic threshold parametrization without requiring any model (re-)training process and therefore is completely standalone.

With the contributions presented, analyzed and concluded in this Thesis, we consider that interesting contributions have been made to the research community that will allow to improve the techniques and results that will be developed in the field of object detection.

8.2 Future work

Based on the results and discussions of this thesis, we propose the following future research lines:

- Part II: Training and Evaluation Frameworks

  - *Chapter 2: Evaluation Framework: Existing and Proposed Datasets and Metrics.* As future work, some methods for autocalibration of cameras can be considered or developed, in order to obtain a common plane between cameras of greater precision. Also a correlation study of the different metrics can be performed in order to analyze the information provided by each, and if some of them are redundant.

  - *Chapter 3: Generation and Evaluation of Synthetic Models for Training People Detectors.* The first approach results are promising and can be improved by generating other more elaborated synthetic image datasets. Observing the masking combination model, some edges are obtained in the area where the wheelchair and the legs join. Smoothing that edges can improve the model. Adding a waist patch can give more realism to the resulting image which can result in a better model. It would be also interesting to test other combinations such as the ones mentioned previously in this chapter: people riding horses, people with shopping carts, etc. Finally, a detection model generated from real images could be completed with this set of generated images in order to improve its detection capacity.

- Part III: Detection Approaches and Applications

- *Chapter 4: Incorporating Wheelchair Users in People Detection.* About the wheelchair users detector, more complex models can be studied, for example considering more model variations. About the combination, we have chosen a simple technique, therefore it could be improved in order to optimize the combination of the different information sources. Also a new model can be trained using both the Smile Lab dataset and the recorded sequences to achieve greater generality. A tracker can be added to the sequence detection to combine the information extracted during the sequence frames giving temporal continuity to the detections. Apart from this, the typical lines of future work for object detection can be applied here. Finally, the improved people detection can be used as a starting point for multiple event detection systems, in scenarios where the presence of wheelchair users is common, such as hospitals, healthcare centers or senior residences.
- *Chapter 5: Automatic vacant parking places management system using multi-camera vehicle detection.* With respect to the combination, we have chosen a simple technique using normalized sigmoid functions, therefore different functions could be studied in order to optimize the combination or fusion of the different information sources. Also a new dataset with more cameras and with different spatial configurations could be recorded to see the behavior of the system in those situations. A tracker can be added to the sequence detection to combine the information extracted during the sequence frames providing temporary continuity to the vehicle detections. Apart from this, current lines of future work for object detection can be applied here, since the detector is the first stage of the system.

- Part IV: Detection Improvements in Multi-camera Scenarios
  - *Chapter 6: Improving multi-camera people detection using contextual information.* As future work, it is proposed to use and combine other detectors to try to improve the system performance and to develop other more elaborated information combination methods, for example by weighing the contribution of each camera to the combined detection as a function of distance between the camera and the object/person, since the closest detections to a camera are usually more accurate and have a greater confidence. In addition, a greater number of cameras can be combined simultaneously, controlling and eliminating the errors introduced by the detections of each camera, using, for example, a majority voting system to decide which detections are kept and which are discarded.
  - *Chapter 7: Enhancing multi-camera people detection by stand-alone automatic parametrization using detection transfer and self-correlation maximization.* More object models for different algorithms can be considered. Also multiple detectors and cameras could

be combined simultaneously, in order to further improve the results. It can also be studied the auto-adaptation of other parameters such as the scale, the detection model, etc. Finally, the techniques of this chapter could be combined with the technique presented in chapter 6, and evaluate its result.

Part VI

Appendixes

Appendix A

Publications

The following publications have been produced in association with this thesis:

- R. Martín-Nieto, J. M. Merchán, Á. García-Martín and J. M. Martínez: "Generation and evaluation of synthetic models for training people detectors," *International Carnahan Conference on Security Technology (ICCST)*, pp. 1-6, 2017.(https://doi.org/10.1109/CCST.2017.8167818):
  - Chapter 3.
- R. Martín-Nieto, Á. García-Martín and J. M. Martínez: "Incorporating Wheelchair Users in People Detection". Under review (08-03-2018)
  - Subsection 2.3.1.
  - Chapter 4.
- R. Martín-Nieto, Á. García-Martín, A. G. Hauptmann, and J. M. Martínez: "Automatic vacant parking places management system using multicamera vehicle detection". Accepted in *IEEE Transactions on Intelligent Transportation Systems* (11-05-2018).
  - Subsections 2.3.2 and 2.6.2.
  - Chapter 5.
- R. Martín-Nieto, A. Miguélez-Sierra, A. García-Martín, J. M. Martínez: Improving multi-camera people detection using contextual information. Under review (31-01-2018).
  - Subsection 2.6.1.
  - Chapter 6.

- R. Martín-Nieto, A. García-Martín, J. M. Martínez, J. C. SanMiguel: "Enhancing multi-camera people detection by stand-alone automatic parametrization using detection transfer and self-correlation maximization". Under review (29-03-2018)
  - Chapter 7.

Appendix B

Logros, conclusiones y trabajo futuro

B.1 Resumen de logros y principales conclusiones

Esta tesis ha abordado diferentes aspectos relacionados con la detección de objetos a tres niveles diferentes: marco de entrenamiento y evaluación (Parte II), procesamiento de detectores y sus aplicaciones (Parte III) y técnicas multi-cámara avanzadas para detección (Parte IV).

Primero, con respecto al marco de entrenamiento y evaluación (Capítulo 2), hemos analizado los trabajos publicados previamente y los conjuntos de datos existentes que cumplian con los requisitos seleccionados, y los hemos completado con la creación de dos conjuntos de datos nuevos: uno considerando gente con distintas apariencias (de pie y usando sillas de ruedas), y otro conjunto de datos grabando un parking real con ejemplos de distinta condiciones de iluminación (de día, de noche, atardecer, etc.) y climáticas (soleado, lluvioso, nublado). Con respecto a las métricas de evaluación, hemos presentado métricas usadas comúnmente para la detección de objetos. Primero, se han formulado las métricas de evaluación 'clásicas', y después hemos adaptado esas métricas para, por un lado considerar una tercera dimensión (profundidad) en los escenarios, y por otro lado evaluar la capacidad de detectar plazas ocupadas o vacías en un parking. Finalmente se han presentado y descrito los algoritmos de detección utilizados en la tesis.

En segundo lugar, la motivación del Capítulo 3 es el estudio de la viabilidad de entrenar un algoritmo de detección de objetos usando un conjunto de datos sintético. En particular, hemos escogido entrenar un modelo de usuario en silla de ruedas utilizando imágenes de sillas de ruedas e imágenes de personas de pie. Se han entrenado tres conjuntos de datos sintéticos distintos para entrenar tres modelos de detección, evaluando qué modelo es el óptimo y finalmente analizando su rendimiento en comparación con un detector de personas en silla de ruedas entrenado con imágenes reales. Con los resultados obtenidos se puede concluir que el rendimiento de los detectores es aceptable, pese a ser peor que el rendimiento obtenido con el detector de sillas de ruedas, entrenado con imágenes reales. Este resultado era de esperar, ya que las imágenes

generadas sintéticamente son diferentes de las imágenes reales, pero pese a ello se ha obtenido un detector capaz de detectar adecuadamente. A cambio de esta pérdida de rendimiento, se ha obtenido un detector funcional sin la necesidad de grabar los objetos reales (en este caso, usuarios de silla de rueda). este método puede ser útil en situaciones en las que no es posible compilar o grabar un conjunto de datos del objeto deseado, o obtenerlo es demasiado caro en términos de tiempo y recursos.

En tercer lugar, tratamos el problema de variabilidad de apariencia para la detección de las distintas clases del mismo objeto de detección semántico (Capítulo 4). Los escenarios típicos de residencias de ancianos son un ejemplo de esta situación problemática. En particular, nuestro objetivo es detectar de forma simultánea tanto personas de pie como usuarios de silla de ruedas. Por ello se presenta una extensión de la detección de personas que permite detectar personas con necesidad de utilizar sillas de ruedas. Hemos entrenado dos modelos de dos algoritmos para la detección de usuarios en silla de ruedas cuyas detecciones se combinan con las detecciones obtenidas por un modelo de detección de personas de pie tradicional, proporcionando generalidad y capacidad de detección complementaria. Este enfoque no solo se puede aplicar al caso de los usuarios de sillas de ruedas, si no que las ideas expuestas aquí se pueden extrapolar a otros escenarios en los que hay individuos con una apariencia diferente del estándar, como por ejemplo personas usando andadores o bastones.

Describiendo una aplicación distinta para los detectores de objetos, el capítulo 5 presenta un sistema multi-cámara para detección de vehículos y su correspondiente mapeo en las plazas de un parking. El sistema se ha diseñado de forma que las cámaras de seguridad puedan ser usadas para el sistema propuesto tras una simple configuración, sin la necesidad de un despliegue nuevo completo de cámaras. El sistema diseñado puede afrontar escenarios más complicados que los considerados en el estado del arte: oclusiones casi totales y cambios climáticos (escenareios nublados, lluvia, nieve, sol...) que limita o reduce su rendimiento. Adicionalmente, se añade la consideración de un escenario multi-cámara, que hasta donde sabemos, no ha sido publicado para este tipo de sistemas.

El primer capítulo de la parte de las técnicas avanzadas multi-cámara para detección (Capítulo 6) presenta un sistema para mejorar la salida de los algoritmos de detección de objetos, transfiriendo y combinando la información obtenida de las distintas cámaras. El sistema añade información de contexto de la escena a un detector de objetos, obteniendo una mejora en el rendimiento de cada cámara de forma independiente. Los resultados obtenidos tras la evaluación del sistema (a nivel de cámaras y a nivel de plano del suelo) confirman la hipótesis inicial, obteniendo mejoras en el rendimiento de detección, combinando información en los cuatro escenarios evaluados. El método propuesto funciona para distinta relación de aspecto (personas de pie y en silla de ruedas) y para cualquier orientación entre las distintas cámaras, gracias a la asunción volumétrica propuesta (un cilindro en el caso de personas). La estimación del cilindro

y la etapa de transferencia de información son completamente paralelizables para cada cámara, por lo que el incremento computacional con respecto a un sistema estándar es únicamente en la etapa de combinación de detecciones.

Continuando el último trabajo descrito, se ha añadido un marco de correlación para automáticamente parametrizar detectores de objetos durante el tiempo de ejecución (Capítulo 7). Esta propuesta aprovecha la correlación entre las detecciones de múltiples cámaras transferidas a una única cámara para determinar el mejor umbral para cada cámara. La aproximación propuesta es capaz de funcionar sobre la salida estándar de detectores del estado del arte, por lo que permite considerar cualquier técnica de detección y modelo de objeto. Este marco perimte automatizar la parametrización del umbral de decisión sin necesidad el proceso de (re-)entrenamiento y por lo tanto es complétamente autónomo.

Con estas contribuciones presentadas, analizadas y concluidas en la tesis, consideramos que se han hecho contribuciones interesasntes a la comunidad investigadora que permitirán mejorar las técnicas y resultados que se desarrollarán en el futuro para el campo de detección de objetos.

B.2 Trabajo futuro

En base a los resultados y discusiones de esta tesis, proponemos las siguientes líneas de trabajo futuro:

- Parte II: Marcos de entrenamiento y evaluación
  - *Capítulo 2: Conjuntos de datos, métricas y algoritmos de detección existentes y propuestos.* Como trabajo futuro, algunos métodos para autocalibración de cámaras pueden ser considerados o desarrollados, con el fin de obtener un plano común entre cámaras con mayor precisión. Además un estudio de correlación de las distintas métricas de evaluación puede ser realizado con el fin de analizar la información aportada y conocer la redundancia entre ellas.
  - *Capítulo 3: Generación y evaluación de modelos sintéticos para el entrenamiento de detectores de personas.* Los primeros resultados evaluados son prometedores y pueden ser mejorados generando conjuntos de imágenes más elaborados. Observando el el modelo de combinación mediante enmascaramiento, algunos bordes se obtienen en el área en el que las piernas y la silla de ruedas se unen. Emborronar esos bordes pueden mejorar el modelo. Añadir un parche de la cintura puede dar más realismo a la imagen resultante, lo que puede resultar en un mejor modelo. También sería interesante probar otras combinaciones de objetos como las mencionadas en el capítulo: jinetes de caballos, personas llevando carros de la compra, etc. Finalmente, un modelo de

detección entrenado con imágenes reales puede ser completado con el conjunto de las imágenes generados con el fin de mejorar su capacidad de detección.

- Parte III: Aproximaciones de detección y aplicaciones

  - *Capítulo 4: Incorporación de usuarios en silla de ruedas en detección de personas.* Con respecto al detector de usuarios de silla de ruedas, se pueden estudiar modelos más complejos, por ejemplo considerando más variaciones del modelo. Con respecto a la combinación, hemos escogido una técnica simple, por lo que puede ser mejorada con el fin de optimizar la combinación de las distintas fuentes de información. También se puede entrenar un modelo de detección utilizando tanto el conjunto de datos de SMILE como las secuencias grabadas, con el fin de alcanzar una mayor generalidad. Se puede añadir un tracker a la secuencia de detección para combinar la información extraida durante los distintos frames, aportando continuidad temporal a las detecciones. Aparte de las líneas mencionadas, las líneas de investigación típicas para detección de objetos pueden ser aplicadas aquí. Finalmente, el detector de personas mejorado puede ser usado como el punto de partida para un sistema de detección de múltiples eventos, en escenarios en los que la presencia de usuarios de silla de ruedas es común, tales como hospitales, centros de salud o residencias de ancianos.
  - *Capítulo 5: Sistema para la gestión automática de plazas vacías mediante detección de vehículos multi-cámara.* Con respecto a la combinación, hemos elegido una técnica simple que utiliza sigmoides normalizadas, por lo que se podrían estudiar diferentes funciones para optimizar la combinación o fusión de las diferentes fuentes de información. También se podría grabar un nuevo conjunto de datos con más cámaras y con diferentes configuraciones espaciales para ver el comportamiento del sistema en esas situaciones. Se puede agregar un sistema de seguimiento sobre las detecciones de la secuencia para combinar la información extraída durante los fotogramas proporcionando continuidad temporal a las detecciones de vehículos. Aparte de esto, las líneas actuales de trabajo futuro para la detección de objetos se pueden aplicar a este trabajo, ya que el detector es la primera etapa del sistema.

- Parte IV: Mejoras de detección en escenarios multicámara

  - *Capítulo 6: Mejora de detección multi-cámara de personas mediante información contextual.* Como trabajo futuro, se propone usar y combinar otros detectores para tratar de mejorar el rendimiento del sistema y desarrollar otros métodos de combinación de información más elaborados, por ejemplo, ponderando la contribución de cada cámara a la detección combinada en función de la distancia entre el la cámara y

el objeto, ya que las detecciones más cercanas a una cámara suelen ser más precisas y tienen una mayor confianza. Además, se puede combinar simultáneamente un mayor número de cámaras, controlando y eliminando los errores introducidos por las detecciones de cada cámara, utilizando, por ejemplo, un sistema de votación por mayoría para decidir qué detecciones se mantienen y cuáles se descartan.

– *Capítulo 7: Mejora de detección multicámara de personas mediante parametrización automática independiente utilizando transferencia de detección y maximización de autocorrelación.* Más modelos de objetos de distintos algoritmos piueden ser considerados. Además múltiples detectores y cámaras pueden combinarse simultáneamente, con el fin de mejorar aún más los resultados. También se puede estudiar la autoadaptación de otros parámetros tales como escala, modelo de detección, etc.

Glossary

AP *Average Precision*

ACF *Aggregated Channel Features (detector)*

AUC *Area Under the Curve*

BB *Bounding Box (also BBox)*

CCH *Context Contrast Histogram*

CDF *Cumulative Distribution Function*

CNN *Convolutional Neural Networks*

DoG *Difference of Gaussian*

DPM *Discriminatively trained Part-based Models (detector)*

FN *False Negative(s)*

FND *False Negative(s) Detection(s)*

FP *False Positive(s)*

FPD *False Positive(s) Detection(s)*

GT Ground Truth

HOG *Histogram of Oriented Gradients*

ITS *Intelligent Transportation Systems*

K-NN *k-Nearest Neighbors*

LDA *Linear Discriminant Analysis*

NN *Neural Network*

OCR *Optical Character Recognition*

PCA *Principal Component Analysis*

PDF *Probability density Function*

PLds *Parking Lot dataset*

PR *Precision*

R-CNN *Region-based Convolutional Neural Networks (detector)*

RE *Recall*

ROI *Region Of Interest*

RPN *Region Proposal Network*

SP *Standing People*

STME *Surface Texture and Microstructure Extraction*

SVM *Support Vector Machine*

TP *True Positive(s)*

TPD *True Positive(s) Detection(s)*

WU *Wheelchair User(s)*

WUds *Wheelchair Users dataset*

YOLO *You Only Look Once (detector)*

Bibliography

H. R. H. Al-Absi, J. D. D. Devaraj, P. Sebastian, and Y. V. Voon. Vision-based automated parking system. In *International Conference on Information Science, Signal Processing and their Applications*, pages 757–760, 2010. [Cited on pages 56, 58, and 62.]

I. P. Alonso, D. F. Llorca, M. A. Sotelo, L. M. Bergasa, P. R. de Toro, J. Nuevo, M. Ocaña, and M. A. G. Garrido. Combination of feature extraction methods for svm pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 8(2):292–307, 2007. [Cited on pages 27 and 28.]

M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *In Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 1–8, 2008. [Cited on pages 20 and 22.]

M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *In Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 1014–1021, 2009. [Cited on page 28.]

E. Auvinet, F. Multon, A. Saint-Arnaud, J. Rousseau, and J. Meunier. Fall detection with multiple cameras: An occlusion-resistant method based on 3-d silhouette vertical distribution. *IEEE Transactions on Information Technology in Biomedicine*, 15(2):290–300, 2011. ISSN 1089-7771. [Cited on page 40.]

H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008. [Cited on page 56.]

Z.-P. Bian, J. Hou, L.-P. Chau, and N. Magnenat-Thalmann. Fall detection based on body part tracking using a depth camera. *IEEE Journal of Biomedical and Health Informatics*, 19(2):430–439, 2015. [Cited on page 40.]

J. Black, T. Ellis, and P. Rosin. Multi view image surveillance and tracking. *Proceedings of the Workshop on Motion and Video Computing*, pages 169–174, 2002. [Cited on page 80.]

K. Blumer, H. R. Halaseh, M. U. Ahsan, H. Dong, and N. Mavridis. *Cost-Effective Single-Camera Multi-Car Parking Monitoring and Vacancy Detection towards Real-World Parking Statistics and Real-Time Reporting*, pages 506–515. Springer Berlin Heidelberg, 2012. [Cited on pages 54 and 55.]

B. L. Bong, K. C. Ting, and K. C. Lai. Integrated approach in the design of car park occupancy information system (coins). *International Journal of Computer Science*, 35(1):1–8, 2008. [Cited on pages 54, 55, 58, and 62.]

D. B. Bong, K. C. Ting, and N. Rajaee. Car-park occupancy information system. *Real-Time Technology And Applications Symposium*, pages 1–4, 2006. [Cited on pages 54 and 55.]

L. C. Chen, J. W. Hsieh, W. R. Lai, C. X. Wu, and S. Y. Chen. Vision-based vehicle surveillance and parking lot management using multiple cameras. In *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 631–634, 2010. [Cited on pages 54 and 55.]

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition*, pages 886–893, 2005. [Cited on pages 23, 28, 41, 44, 56, and 60.]

J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *International Conference on Machine Learning*, pages 233–240, 2006. [Cited on page 20.]

F. de Chaumont, B. Marhic, L. Delahoche, and C. Cauchois. Generic method for recognition of a wheelchair, even with a low resolution-effective sensor. In *International Conference on Industrial Technology*, pages 56–60, 2004. [Cited on page 29.]

P. Dollar, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *Transactions on IEEE Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014. [Cited on pages 23, 60, and 97.]

P. Dollár, R. Appel, and W. Kienzle. Crosstalk cascades for frame-rate pedestrian detection. In *European Conference on Computer Vision*, pages 645–659, 2012a. [Cited on page 28.]

P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *Transactions on IEEE Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012b. [Cited on page 28.]

M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *Transactions on IEEE Pattern Analysis and Machine Intelligence*, 31(12):2179–2195, 2009. [Cited on pages 27 and 28.]

T. Fabian. An algorithm for parking lot occupation detection. In *Computer Information Systems and Industrial Management Applications*, pages 165–170, 2008. [Cited on pages 53 and 54.]

P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/, 2010a. [Cited on pages 41 and 42.]

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Transactions on IEEE Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010b. [Cited on pages 23, 27, 28, 31, 40, 41, 60, 69, 84, and 97.]

. García-Martín, A. Cavallaro, and J. M. Martínez. People-background segmentation with unequal error cost. In *IEEE International Conference on Image Processing*, pages 157–160, 2012. [Cited on page 27.]

A. García-Martín and J. M. Martínez. Robust real time moving people detection in surveillance scenarios. In *International Conference on Advanced Video and Signal based Surveillance*, pages 241–247, 2010. [Cited on page 28.]

A. García-Martín and J. M. Martínez. Post-processing approaches for improving people detection performance. *Computer Vision and Image Understanding*, 133:76 – 89, 2015a. [Cited on page 44.]

A. García-Martín and J. M. Martínez. People detection in surveillance: classification and evaluation. *IET Computer Vision*, 9(5):779–788, 2015b. [Cited on pages 26, 27, and 80.]

A. García-Martín and J. C. SanMiguel. Adaptive people detection based on cross-correlation maximization. In *IEEE International Conference on Image Processing*, pages 3385–3389, 2017. [Cited on page 96.]

D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *Transactions on IEEE Pattern Analysis and Machine Intelligence*, 32(7): 1239–1258, 2010. [Cited on page 28.]

R. B. Girshick. Fast R-CNN. *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference*, pages 1440–1448, 2015. [Cited on pages 23, 41, 57, and 60.]

R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference*, pages 580–587, 2013. [Cited on pages 23, 41, 57, and 60.]

R. B. Girshick, F. N. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference*, 2014. [Cited on page 28.]

R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003. [Cited on page 59.]

I. A.-B. Hilal Al-Kharusi. Intelligent parking management system based on image processing. *World Journal of Engineering and Technology*, 2(2):55–67, 2014. [Cited on pages 54, 55, 58, and 62.]

D. Hosotani, I. Yoda, and K. Sakaue. Wheelchair recognition by using stereo vision and histogram of oriented gradients (hog) in real environments. In *Workshop on Applications of Computer Vision*, pages 1–6, 2009. [Cited on pages 29 and 39.]

W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(3): 334–352, 2004. [Cited on page 27.]

C. C. Huang and H. T. Vu. A multi-layer discriminative framework for parking space detection. In *International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2015. [Cited on pages 56, 57, and 58.]

C. C. Huang and S. J. Wang. A hierarchical bayesian generation framework for vacant parking space detection. *Transactions on Circuits and Systems for Video Technology*, 20(12):1770–1785, 2010. [Cited on pages 53, 54, and 55.]

C.-C. Huang, S.-J. Wang, Y.-J. Change, and T. Chen. A bayesian hierarchical detection framework for parking space detection. In *International Conference on Acoustics, Speech and Signal Processing*, pages 2097–2100, 2008. [Cited on page 56.]

C. C. Huang, Y.-S. Dai, and S. J. Wang. A surface-based vacant space detection for an intelligent parking lot. In *International Conference on ITS Telecommunications*, pages 284–288, 2012. [Cited on pages 56 and 57.]

C. C. Huang, Y. S. Tai, and S. J. Wang. Vacant parking space detection based on plane-based bayesian hierarchical framework. *Transactions on Circuits and Systems for Video Technology*, 23(9):1598–1610, 2013a. [Cited on pages 56, 57, and 58.]

C. C. Huang, H. T. Vu, and Y. R. Chen. A multiclass boosting approach for integrating weak classifiers in parking space detection. In *International Conference on Consumer Electronics*, pages 314–315, 2015. [Cited on pages 56, 57, and 58.]

C.-R. Huang, P.-C. Chung, K.-W. Lin, and S.-C. Tseng. Wheelchair detection using cascaded decision tree. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):292–300, 2010. [Cited on pages 14, 29, 45, 47, and 48.]

P.-J. Huang and D. yu Chen. Robust wheelchair pedestrian detection using sparse representation. In *Visual Communications and Image Processing*, pages 1–5, 2012. [Cited on page 29.]

Y.-X. Huang, S.-P. Hsu, C.-C. Yu, Y.-N. Chung, and C.-T. Lin. Applying image technology to detect and track the wheelchair patient safety. In *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, pages 2333–2415, 2013b. [Cited on page 28.]

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. [Cited on page 41.]

P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, and N. Papanikolopoulos. Estimating pedestrian counts in groups. *Computer Vision and Image Understanding*, 110:43–59, 2008. [Cited on pages 27, 81, and 95.]

K. Kim and L. S. Davis. *Multi-camera Tracking and Segmentation of Occluded People on Ground Plane Using Search-Guided Particle Filtering*, pages 98–109. Springer Berlin Heidelberg, 2006. [Cited on page 80.]

C.-H. Lee, M.-G. Wen, C.-C. Han, and D.-C. Kou. An automatic monitoring approach for unsupervised parking lots in outdoors. In *IEEE International Carnahan Conference on Security Technology*, pages 271–274, 2005. [Cited on pages 54 and 55.]

B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. *Proceedings of the IEEE Computer Vision and Pattern Recognition conference*, pages 878–885, 2005. [Cited on pages 17, 20, and 82.]

B. Leibe, K. Schindler, and L. V. Gool. Coupled detection and trajectory estimation for multi-object tracking. *IEEE International Conference on Computer Vision*, pages 1–8, 2007. [Cited on page 27.]

B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77:259–289, 2008. [Cited on pages 20, 22, 27, and 28.]

S. F. Lin, Y. Y. Chen, and S. C. Liu. A vision-based parking lot management system. In *International Conference on Systems, Man and Cybernetics*, volume 4, pages 2897–2902, 2006. [Cited on pages 54, 55, 58, and 62.]

J. Liu, M. Mohandes, and M. Deriche. A multi-classifier image based vacant parking detection system. In *International Conference on Electronics, Circuits, and Systems*, pages 933–936, 2013. [Cited on pages 54 and 55.]

R. Martín-Nieto, . García-Martín, A. G. Hauptmann, and J. M. Martínez. Automatic vacant parking places management system using multicamera vehicle detection. *IEEE Trans. on Intelligent TransportatUnder Review (In press)*, pages 1–12, 2017. [Cited on page 53.]

R. Martín-Nieto, . García-Martín, and J. M. Martínez. Incorporating wheelchair users in people detection. *Under Review*, 2018a. [Cited on page 39.]

R. Martín-Nieto, A. García-Martín, J. M. Martínez, and J. C. SanMiguel. Enhancing multicamera people detection by stand-alone automatic parametrization using detection transfer and self-correlation maximization. *Under Review*, 2018b. [Cited on page 93.]

R. Martín-Nieto, A. Miguélez-Sierra, A. García-Martín, and J. M. Martínez. Improving multicamera people detection using contextual information. *Under Review*, 2018c. [Cited on page 79.]

I. Masmoudi, A. Wali, A. Jamoussi, and A. M. Alimi. Vision based system for vacant parking lot detection: VPLD. In *International Conference on Computer Vision Theory and Applications*, volume 2, pages 526–533, 2014. [Cited on pages 54 and 56.]

A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *Transactions on the IEEE Pattern Analysis and Machine Intelligence*, 36:58–72, 2014. [Cited on pages 12 and 96.]

A. Myles, N. da Vitoria Lobo, and M. Shah. Wheelchair detection in a calibrated environment. In *Asian Conference on Computer Vision*, pages 1–7, 2002. [Cited on page 28.]

W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, Y. Xiong, C. Qian, Z. Zhu, R. Wang, C. C. Loy, X. Wang, and X. Tang. DeepID-Net: multi-stage and deformable deep convolutional neural networks for object detection. *In proceedings of the IEEE Computer Vision and Pattern Recognition*, 2014. [Cited on page 28.]

. G.-M. R. Martín-Nieto, J. M. Merchán and J. M. Martínez. Generation and evaluation of synthetic models for training people detectors. In *International Carnahan Conference on Security Technology*, pages 1–6, 2017. [Cited on page 25.]

J. Redmon and A. Farhadi. YOLO9000: Better, Faster, Stronger. In *IEEE Conference on Computer Vision and Pattern Recognition conference*, pages 6517–6525, 2017. [Cited on pages 23 and 97.]

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. [Cited on page 23.]

S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *proceedings of the IEEE Computer Vision and Pattern Recognition conference*, pages 91–99, 2015. [Cited on pages 23, 28, 40, 41, 42, 57, 59, 60, 69, and 97.]

T. T. Santos and C. H. Morimoto. People detection under occlusion in multiple camera views. *Proceedings of Brazilian Symposium on Computer Graphics and Image Processing*, pages 53–60, 2008. [Cited on page 80.]

T. T. Santos and C. H. Morimoto. Multiple camera people detection and tracking using support integration. *Pattern Recognition Letters*, 32(1):47–55, 2011. [Cited on page 80.]

R. J. L. Sastre, P. G. Jimenez, F. J. Acevedo, and S. M. Bascon. Computer algebra algorithms applied to computer vision in a parking management system. In *International Symposium on Industrial Electronics*, pages 1675–1680, 2007. [Cited on pages 56, 58, and 62.]

D. Simonnet, S. Velastin, E. Turkbeyler, and J. Orwell. Backgroundless detection of pedestrians in cluttered conditions based on monocular images: a review. *IET Computer Vision*, 6(6):540–550, 2012. [Cited on page 28.]

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 2014. [Cited on pages 43 and 60.]

N. True. Vacant parking space detection in static images. *Projects in Vision & Learning*, pages 1–5, 2007. [Cited on pages 56, 58, and 62.]

R. Y. Tsai. An efficient and accurate camera calibration technique for 3d machine vision. *Proceedings of the IEEE Computer Vision and Pattern Recognition conference*, pages 364–374, 1986. [Cited on pages 12, 84, and 96.]

M. Tschentscher, C. Koch, M. König, J. Salmen, and M. Schlipsing. Scalable real-time parking lot classification: An evaluation of image features and supervised learning algorithms. In *International Joint Conference on Neural Networks*, pages 1–8, 2015. [Cited on pages 56 and 57.]

M. Valera and S. A. Velastin. Intelligent distributed surveillance systems: a review. *Visual Image Signal Processing*, 152(2):192–204, 2005. [Cited on page 27.]

P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2): 137–154, 2004. [Cited on page 28.]

X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, 34(1):3–19, 2013. [Cited on page 80.]

X. Wang and A. R. Hanson. Parking lot analysis and visualization from aerial images. In *IEEE Workshop on Applications of Computer Vision*, pages 36–41, 1998. [Cited on pages 54 and 55.]

C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *Computer Vision and Pattern Recognition*, pages 794–801, 2009. [Cited on pages 20 and 22.]

C.-W. Wu, C.-D. Liu, and P.-C. Chung. Assistance instruments detection using geometry constrained knowledge for health care centers. In *International Conference on Future Information Technology*, pages 1–5, 2010. [Cited on page 28.]

Q. Wu, C. Huang, S. y. Wang, W. c. Chiu, and T. Chen. Robust parking space detection considering inter-space correlation. In *International Conference on Multimedia and Expo*, pages 659–662, 2007. [Cited on pages 56 and 58.]

H. Xie, Q. Wu, B. Chen, Y. Chen, and S. Hong. Vehicle detection in open parks using a convolutional neural network. In *International Conference on Intelligent Systems Design and Engineering Applications*, pages 927–930, 2015. [Cited on pages 57, 58, and 60.]

K. Yamada and M. Mizuno. A vehicle parking detection method using image segmentation. *Electronics and Communications in Japan*, 84(10):25–34, 2001. [Cited on pages 54 and 55.]

C.-A. Yang and P.-C. Chung. Recovery of 3-d location and orientation of a wheelchair in a calibrated environment by using single perspective geometry. In *Region 10 Conference*, pages 1–4, 2007. [Cited on page 28.]